High-level visual object representation in juvenile and adult primates

by

Darren Seibert

Submitted to the Department of Brain and Cognitive Sciences
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2018

Signature redacted

Author . . . .
Department of Brain and Cognitive Sciences
August 7, 2018

Signature redacted

Certified by . . . .
James J. DiCarlo
Professor of Neuroscience
Head, Department of Brain and Cognitive Sciences
Thesis Supervisor

Signature redacted

Accepted by . . . .
Matthew A. Wilson
Sherman Fairchild Professor of Neuroscience and Picower Scholar
Director of Graduate Education for Brain and Cognitive Sciences

High-level visual object representation in juvenile and adult primates

by

Darren Seibert

Submitted to the Department of Brain and Cognitive Sciences
on August 7, 2018, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Despite being reflexive, primate view invariant object recognition is a complex computational task. These computations are thought to reside in the ventral visual stream, specifically culminating in inferior temporal (IT) cortex. Recent research in machine learning has made great progress in modeling primate ventral visual stream computations. While the end result of current machine learning approaches produces models that are highly predictive of the adult state of the ventral stream, the learning approaches themselves are not biologically plausible, requiring tens of thousands to millions of human-labeled training points. Understanding primate visual development is therefore not only interesting from the perspective of neuroscience, but also has practical value in building more robust learning algorithms capable of functioning in domains where large amounts of human-labeled training information may be difficult or impossible to create. Better learning algorithms may also produce agents capable of adapting and behaving in the world not unlike humans. This thesis first describes work on predicting visual responses across the human ventral stream using convolutional neural networks (CNNs). We then describe a set of natural image statistics automatically incorporated into high-performing CNNs from supervised training—it is possible primate development incorporates these or similar natural image statistics into its synaptic strengths. Finally, we describe the first-large scale characterization of IT in 19-32 week old macaques. While we find longer response latencies in these younger animals, we do not find any differences in representation between adults and juveniles suggesting that, at 19-32 weeks of age, IT already supports robust object recognition consistent with adults. Our data provide an upper limit on the amount of training data needed to reach adult-level performance—approximately 2,800 hours of waking visual experience.

Thesis Supervisor: James J. DiCarlo
Title: Professor of Neuroscience
Head, Department of Brain and Cognitive Sciences

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Primates excel at visual object recognition and are capable of recognizing thousands of objects in countless variations, positions, and lighting conditions in tenths of a second. The ease at which we recognize objects is not a trivial feat—approximately one third of cortex is involved in some form of visual processing and, until recently, creating algorithms to solve this task eluded multiple fields of study. Our ability to recognize visual objects is thought to occur in the ventral visual stream, whereby information flows in a largely feed-forward fashion from lower to higher visual areas where representations grow increasingly more categorical. Much work has shed light on lower ventral visual areas [1], but the neural representations in higher ventral cortical areas have until recently proven extremely difficult to characterize [2, 3]. In the last few years, however, there has been significant progress toward better understanding of higher ventral visual cortex, combining advances in neurophysiological measurement techniques [4, 5] with leaps in computer vision and computational modeling technology [6, 7].

Recent progress in neurophysiological measurement techniques have demonstrated convincingly that IT cortex contains a robust representation of object category and identity information [4, 5, 8, 9], while advances in computational modeling have provided quantitatively predictive models for adult V4 and IT neural response patterns [10, 11].

While these advances provide effective descriptions of the "adult state" of visual representation [10], the dynamic learning rules by which visual areas develop remain unknown. Based on converging evidence from fMRI [12, 13], unsupervised IT learning [14], and powerful neural network learning procedures [6, 15, 16, 17, 18], we hypothesize that visual cortex is refined postnatally, incorporating a myriad of image statistics [14, 19, 20, 21, 22]. However, it is far from clear exactly what these learning rules are, especially in higher visual cortical areas such as IT. Even the magnitude and nature of developmental changes, relative to hard-wired pre-natal cortical structure, remain unclear. Indeed, visual sensing, occurs even in some single-celled bacteria and may have played an important evolutionary role in shaping the primate brain.

1.1 Developmental electrophysiology

Anatomical and electrophysiological developmental studies of vision are often suggestive of early maturation with some dependence on visual experience. These studies are not typically motivated from a computational or behavioral perspective. As a consequence, the phenomena tested often have unclear bearing on how they might influence the flow of information through visual areas and how developmental changes may influence downstream decoders. For example, development of orientation selectivity occurs without visual experience [23] and can be predicted by thalamic inputs [24], though visual experience can change orientation selectivity in V1 [25], but the significance of these effects in improving the performance of downstream decoders is not contextualized. Furthermore, visual experience plays a role in maintaining and sharpening selectivity, in addition to pruning connections from non-visually responsive neurons [26]. There is evidence that V1 ocular dominance is fully adult-like as early as 6 weeks, with clear ocular dominance as early as 3 weeks [27]. Generally, studies of early visual areas suggest that experience may play more of a role in maintaining these anatomical and single-unit response properties rather than in creating them [28]. Cortical projections to IT appear adult-like by 7 to 18 weeks [29] and projections from IT to parahippocampus and perirhinal cortex appear adult-like at one week. Using cartoon-like face stimuli, awake infant IT appears adult-like at the earliest tested age of one month [30].

These studies do not show that invariant object recognition in higher visual cortex is present from birth or even soon after. One problem is that much of the visual development literature focuses on early visual areas and/or rodent models. However, mice and rats differ substantially in visual function from human and higher primate vision at all levels of the visual hierarchy [31, 32], having substantially lower visual acuity [33], no color vision, lacking clear ocular dominance and columnar organization [34, 35], and having "higher" visual areas that may only differentially respond to spatial or temporal frequencies [31] (as opposed to the high-level object statistics encoded by primate higher visual cortex). Moreover, most of the existing juvenile macaque studies (and all such chronic studies) were performed in anesthetized animals, which in general showed significantly lower numbers of visually driven cells in higher visual cortex than adult animals when anesthetized [30]. However, the few studies in awake juveniles show that the general level of visually driven cells does not appear to be lower than in adults [30]. This suggests the infant visual cortex is at a gross physiological level more sensitive to the effects of anesthesia than adults, which complicates the interpretation of the anesthetized recordings in regards to development of underlying visual representation. Additionally, simplistic stimuli used in many existing studies (ex. no position/pose/size variation, cartoon images, no backgrounds), do not expose several important axes of variation that make higher-level object recognition challenging [36]. Thus, these studies may not be sensitive to the key computational machinery that is present in adult IT.

1.2 Development of object and category recognition

While electrophysiological studies of early visual areas are often suggestive of maturation on the scale of post-natal weeks, studies of higher visual cortex and its representation of faces and objects are often suggestive of slower maturation. Human functional imaging studies have demonstrated that while face selective regions are present at the earliest tested ages (3-8 months—corresponding to approximately 1-2 month old macaques) [37], they do not have completely adult-like responses until adolescence [12] and their representational similarities may be different from adults [37]. Preliminary juvenile macaque functional imaging results indicate that category selective areas are not observable until around 40 post-natal weeks and that face selectivity does not fully emerge until 2 years [38]. Learning experiments in adult macaques also show significant residual neural plasticity in response to visual experience [21, 20] and suggest that face experience may be required for development of face patches [39]. Macaque psychophysical studies often point towards gradual emergence of performance on a variety of tasks [40, 28], though these studies cannot localize where neuronal development occurs, since most or all of the behavioral change might be due to development in downstream decision-making or motor action circuitry, rather than in the visual representation itself [40].

An emerging hypothesis in the field that attempts to explain all of these data is that later visual areas both take longer to develop and are more plastic [41, 42]. Precisely this sort of developmental trajectory can be observed during the training a convolutional neural network using supervised error backpropagation [43], where gross orientation selectivity emerges in first-layer model filters early on (Figure 1-1b), consistent with the observations in V1 early-development studies [27, 26]. However, achieving ecologically-relevant levels of object recognition performance takes an order of magnitude more time (Figure 1-1b).

1.3 Models of learning

While a number of learning rules have been proposed in the literature, it is not until recently that any learning rule—bio-plausible or not—has been capable of generating models approaching human levels of object recognition performance (Figure 1-1a). As is shown in Figure 1-1a, higher levels of object recognition performance are generally predictive of higher neural predictivity. We further this line of work in Chapter 2. Although it is likely that the actual learning procedure used in primate development shares properties of previously proposed unsupervised learning rules—

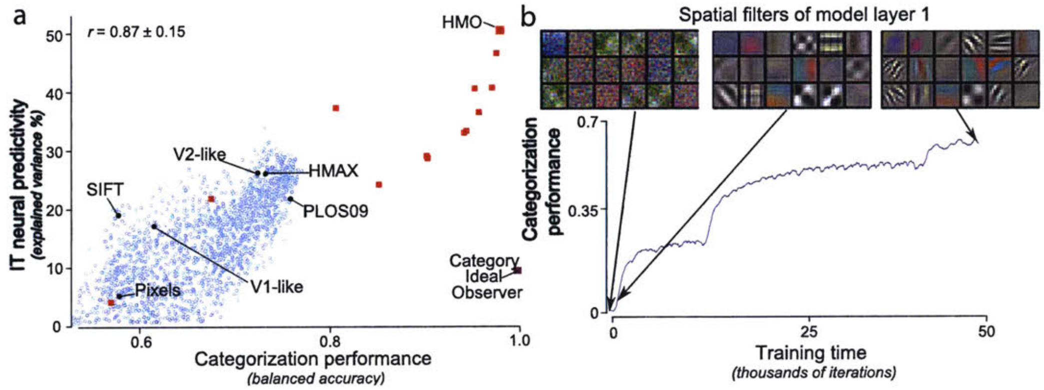

Figure 1-1: Modeling the ventral stream
a. Recent modeling results have made significant strides in capturing neural response patterns in higher ventral cortex, including IT and V4 cortex (Figure adapted from [10]). This work has also shown that object categorization performance is strongly correlated to neural predictivity across a variety of models exhibiting a wide range of performance levels. Thus, tracking performance levels in models of this kind is likely to be a good proxy for neural similarity to the adult IT representation. **b.** Performance vs. time for a convolutional neural network model of inferior temporal cortex, trained via supervised error back-propagation on a categorization task. Mimicking the known properties of V1 development, orientation selectivity emerges very early on in model training (see filter insets). In contrast, categorization performance takes much longer to fully emerge. Performance is highly sensitive to parameter values — changes on the order of 0.1% of the values can cause large changes to performance [43], but will leave orientation tuning apparently unchanged. For this reason, gross developmental features of lower visual areas are likely not indicative of critical but subtle changes in higher visual areas.

for example slowness [22], sparseness [44, 45], and Hebbian-like features [46]—these ideas are not in themselves seemingly sufficient to build representations capable of supporting invariant object recognition. Likely, a more general optimization function is being employed which implicitly results in these properties. As an example: recent modeling work [47, 48, 49, 50, 51] has proposed a number of unsupervised approaches to training neural networks revolving around the idea of generative networks. These networks have been used to predict pixels [50], images [52], and sequences of frames of natural photographs and videos [49] and provide an approach to training models which does not rely on human-labeled semantic category labels. In some cases, increased performance on these tasks indirectly results in networks which can also robustly recognize objects by training linear classifiers on higher-level network output layers.

While the previously discussed pixel, image, and frame prediction tasks appear promising, it is conceivable that none of these are the exact task used to drive ventral stream development. If true, the general task-based optimization approach still may prove to be useful in shaping how we view unsupervised learning. The exact task used by primate vision is likely one whose performance necessitates robust object recognition abilities and whose performance can be assessed with minimal external information. For instance, perfect frame prediction far enough into the future implicitly requires not only object recognition ability, but also information on spatial location, pose, 3D structure, among other latent variables. Additionally, assessing frame prediction performance requires no external supervisory signals and in theory could be implemented entirely by local layer-wise update rules (chained backpropogation, for instance).

Identifying a biologically plausible but computationally effective learning rule for visual development would be of great conceptual interest in systems neuroscience. A long-standing hypothesis is that a few powerful general learning rules are operative in cortical development [53, 46, 45, 16]. Understanding the rules governing visual development we hypothesize will generalize or at least be a step towards having a more universal learning rule. Insight into the learning procedures used by cortex would allow the creation of algorithms capable of adapting and integrating with the natural world in similar ways as humans and will also be beneficial for medicinal goals. In the biomedical domain, a key objective is the treatment of developmental neurological disorders (eg. amblyopia, autism), many of which implicate abnormalities in cortical learning rules generally and visual learning in specific. For example, autistic patients exhibit behavioral deficits in face perception and reduced activation in visual face processing areas [54, 55], a divergence that occurs on the same developmental time scale on which fully adult face perception emerges [12]. Measuring and modeling neurotypical development in higher visual cortex will help provide insight into neural correlates of developmental abnormalities. For AI purposes, an understanding of neuronal learning will aid in the development of more data-efficient algorithms that are capable of training neuronal networks in task domains where limited supervision is available, such as medical diagnostics where millions of labeled training examples are difficult if not impossible to obtain.

1.4 Organization of thesis

This thesis contains three major components. First, we have used high-performing neural networks to predict human functional brain imaging responses. We extend prior work by demonstrating how neural predictivity of models increase as they are trained on an object recognition task—previously it was demonstrated that high performing architectures were better predictors of IT responses [10]. Second, we have observed a set of statistics incorporated by high performing neural networks as a consequence of supervised object recognition training. We test the sufficiency of these statistics alone in producing high-performing models and find that they are able to sustain some, but not all, of the performance indicating that additional statistical information is learned than what we have described.

Third and finally, we describe an experiment in which we have chronically implanted electrode arrays to record neuronal responses to thousands of image stimuli in the IT cortex of awake behaving juvenile macaques. These data represent a snapshot of the developing primate visual representation, and serve as a key counterpart to existing measurements using the same stimuli, including electrophysiology in adult macaques. By comparing juvenile and adult neuronal responses at both single site and population levels, we have obtained an unprecedentedly large-scale and detailed picture of the neural correlates of high-level visual development. These data were collected with two primary goals: (1) test several broad-scale hypotheses predicting the time-course and extent of primate postnatal visual development at approximately 25 weeks of age (see Figure 1-2), (2) provided we observe representational differences, test intermediate states of developmental models (see Figure 1-3).

We have benchmarked the juvenile against adult animals via a battery of performance and representational dissimilarity metrics and find that juveniles at 25 weeks old already contain an IT representation capable of robustly recognizing objects. The one difference we do find between juveniles and adults is a visual onset latency, the significance of which this might or might not have to behavior remains to be understood.

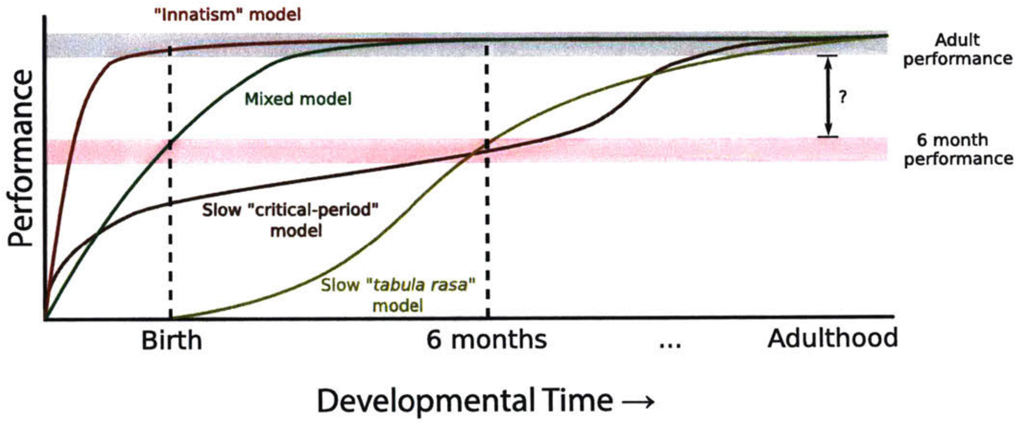

Figure 1-2: Constraining neural developmental time-courses
There are multiple uncertainties regarding primate visual development and, consequently, multiple possible broad-scale time-courses for its maturation. Here, we have sought to address if, and how, the 6 month IT representation differs from adults (using categorization performance and other measures)—as represented by the separation of the red and gray horizontal lines. While recording at 6 months does not give us direct insight into the initial state of IT at birth, it allows us to rule out some possibilities—such as slower maturation (depicted as the slow "critical-period" and slow "tabula rasa" models in the figure) in the case that the juvenile representation is indistinguishable from the adult representation, or, conversely, ruling out faster developmental trajectories (depicted as the "innatism" and "mixed" models in the figure). The mixed model represents an intermediate between the extreme "innatism" (fully developed representation at birth) and slow "tabula rasa" model.

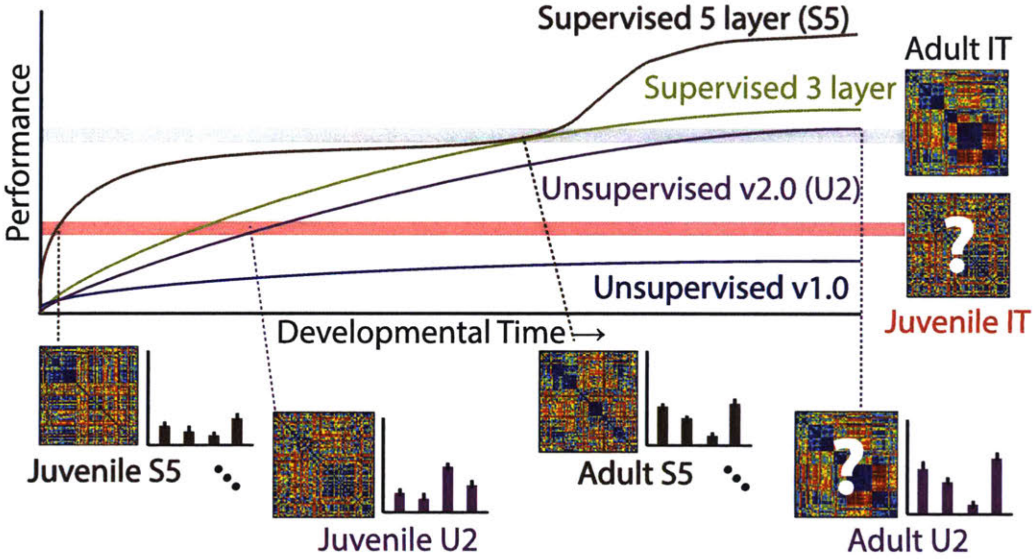

Figure 1-3: Constraining intermediate model representations

We propose that intermediate (non-developed) IT representations can constrain intermediate model representations (and therefore models of primate visual development) by first calibrating the object recognition performance of the model and neural representations. The calibration provides a means to estimate the correspondence between model training and developmental time in weeks because most model training procedures do not explicitly map onto time (many are based on performing gradient descent with some form of input training data). Calibration simply consists of mapping two points in time: (1) the time at which the model reaches the juvenile recognition performance, and (2) the time at which the model reaches adult recognition performance. These two time points can then be used as snapshots in which other, more detailed, metrics can be applied (such as with representational dissimilarity matrices, as shown in the figure [56]). Some models never reach adult performance levels and are automatically ruled out. However, these models could still be compared to intermediate IT representations and conceivably could predict responses better than models which do reach adult recognition levels.

Chapter 2

Modeling the emergence of object recognition in the human ventral stream

Human visual object recognition is subserved by a multitude of cortical areas. To make sense of this system, one line of research focused on response properties of primary visual cortex neurons and developed theoretical models of a set of canonical computations such as convolution, thresholding, exponentiating and normalization that could be hierarchically repeated to give rise to more complex representations. Another line or research focused on response properties of high-level visual cortex and linked these to semantic categories useful for object recognition. Here, we hypothesized that the panoply of visual representations in the human ventral stream may be understood as emergent properties of a system constrained both by simple canonical computations and by top-level, object recognition functionality in a single unified framework [10, 57, 58]. We built a deep convolutional neural network model optimized for object recognition and compared representations at various model levels using representational similarity analysis to human functional imaging responses elicited from viewing hundreds of image stimuli[1]. Neural network layers developed

[1] Work done in collaboration with Daniel Yamins, Diego Ardila, Ha Hong, James DiCarlo, and Justin Gardner

representations that corresponded in a hierarchical consistent fashion to visual areas from V1 to LOC. This correspondence increased with optimization of the model's recognition performance. These findings support a unified view of the ventral stream in which representations from the earliest to the latest stages can be understood as being built from basic computations inspired by modeling of early visual cortex shaped by optimization for high-level object-based performance constraints.

2.1 Significance

Prior work has taken two complimentary approaches to understanding the cortical processes underlying our ability to visually recognize objects. One approach identified canonical computations from primary visual cortex that could be hierarchically repeated and give rise to complex representations. Another approach linked later visual area responses to semantic categories useful for object recognition. Here we combined both approaches by optimizing a deep convolution neural network based on canonical computations to preform object recognition. We found that this network developed hierarchically similar response properties to those of visual areas we measured using functional imaging. Thus, we show that object-based performance optimization results in predictive models that not only share similarity with late visual areas, but also intermediate and early visual areas.

2.2 Introduction

Human cortex contains numerous areas with topographic representations of the visual world [59, 60]. What does each one of these cortical areas ***do***? At least two major divergent approaches to this general question have been taken to understand areas in the ventral visual pathway which is thought to be involved in object vision and perception [61, 62].

One approach, exemplified by research beginning with the primary visual cortex in cats [63] and monkeys [64], has been to examine the visual response properties of neurons and ask mechanistic questions about how properties such as orientation selectivity in simple cells and invariance to position in complex cells are created by neural circuitry [65]. This approach has led to computational models of visual cortical processing in which receptive fields are described as linear weightings [66] of inputs from neurons with center-surround receptive fields [67, 68]. As this linear weighting of visual inputs is performed by neurons with similar filtering properties tiled across the visual field, this stage of processing is akin to a convolution of a filter with visual input. Linear receptive fields are followed by output non-linearites such as thresholding and exponentiation [69, 70, 71] and divisive contrast normalization [69]. These basic computations are proposed to be canonical [72] such that repeating them in a hierarchical fashion [73, 74, 75] may recapitulate computations performed by visual areas along the visual pathways.

A second approach, exemplified particularly by research in humans [76, 77] and monkeys [78, 79, 80, 81, 4] has started largely by asking about whether high-level features of visual scenes such as the presence of objects, faces, places and other identifiable semantic categories are represented in temporal cortex. Links between these representations and perception, for example with faces, are bolstered by similarities between the perceptual phenomenology [82] and representations in ventral cortex [83]. Moreover, causal evidence in the form of lesion [84] and stimulation evidence links high-level representations in the ventral visual stream in both monkeys [85, 86] and humans [87] to perception.

Here we asked if a combination of these two approaches may help explain the nature of response properties not just of early and late areas, but for the full hierarchy of areas in human ventral visual cortex. We used a deep convolutional neural network model [6, 10] whose basic operations were inspired from the canonical computations derived from early visual cortex such as convolution, threshold non-linearities, non-linear pooling and normalization. We also constrained the network to develop high-level representations of object features, by training the network to perform well on invariant object recognition. Previous work has shown that these network models develop representational similarity to V4 and IT in monkey [10, 57] and humans [57, 58]. We capitalized on the ability to measure responses in multiple topographically and functionally localized cortical areas of the human using BOLD imaging to see if this framework could be extended to the whole ventral stream from earliest cortical stages to later ventral areas. While intermediate visual areas such as V2 might be expected to have some kind of intermediate representation between V1 and later stages of the visual system (there are many possible such representations), our model was not explicitly trained to fit V2 responses and therefore was not guaranteed to show any correspondence. Nonetheless, we found representational similarity between the neural network and the human visual system in a hierarchical consistent fashion.

2.3 Materials and Methods

2.3.1 Human subjects

Seven subjects (1 female, ages 22-38) participated. Subjects provided written and oral informed consent before each session and all procedures were approved by the RIKEN Function MRI Safety and Ethics Committee. All subjects underwent at least four imaging sessions (anatomical, retinotopy, category localizer and main experiment). Similar to other studies [88, 89, 90], our analyses required consistent responses to hundreds of image stimuli over many scanning sessions from each subject. Therefore, from the original cohort of subjects, we selected the two which had the highest mean split-half reliability in V1 (see the Stimulus response section) to complete a full data set (at least 9 sessions each consisting of approximately 10 8-minute scans of the main experiment). Of the two pre-screened subjects chosen to complete the full dataset, one was an author. This pre-screening procedure was designed to select subjects based on the overall reliability of data without introducing bias for what representations subjects exhibit. We note that because of the design decision to collect a large data set from a small number of subjects, the results presented here are generalizable only if visual representations in the ventral visual areas across individuals is similar - a notion that is supported by a great deal of literature both within and across species

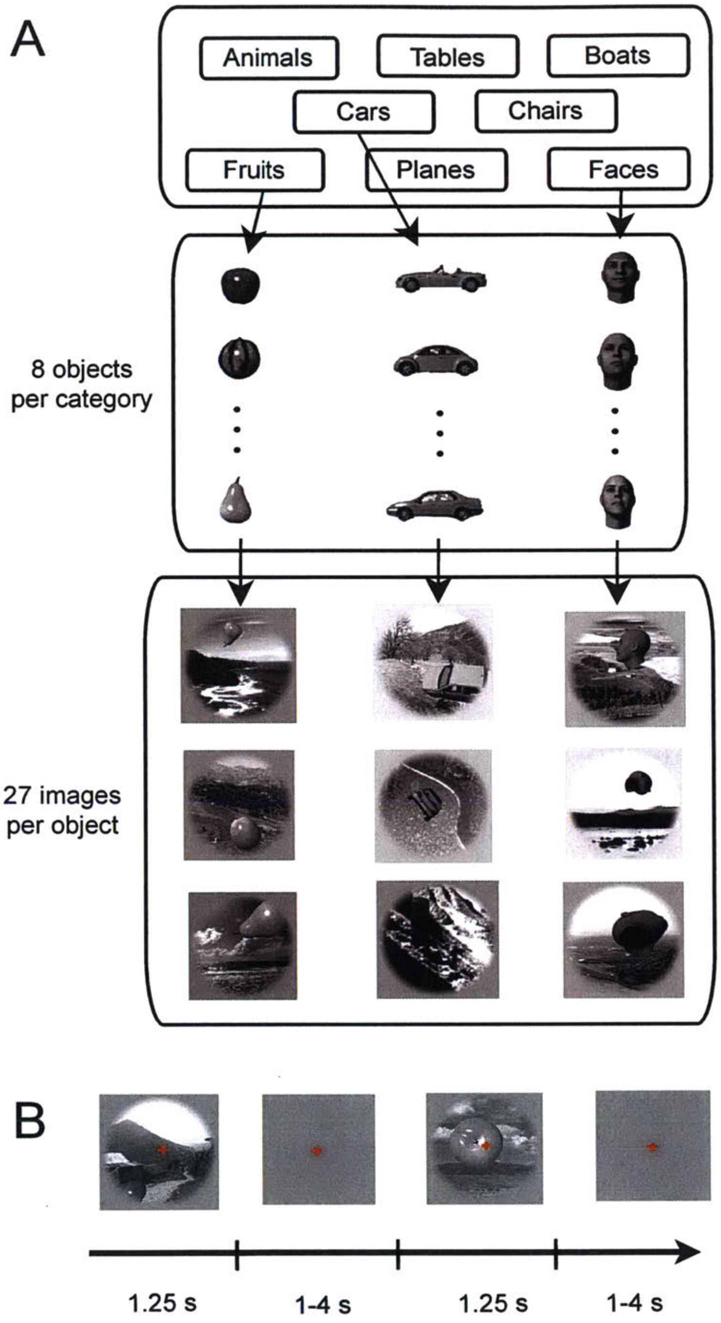

Figure 2-1: Experimental design

Task and stimuli. A, Stimuli contained 8 objects chosen from 8 categories. Each object appeared in 27 or 28 images in random positions, scales, orientations, and on random backgrounds. **B,** Images were shown for 1.25 s followed by a random delay of 1-4 s. Subjects maintained central fixation and performed a discrimination task on the fixation cross.

of primates [56, 10, 57, 88, 89, 90, 58].

2.3.2 Stimuli

We presented 1785 gray-scale images of objects a median of six times across multiple sessions to each subject. Objects were drawn from 8 categories (animals, tables, boats, cars, chairs, fruits, planes, and faces) containing 8 exemplars. Each object was shown from 27 or 28 different viewpoints against a random natural background (circular vignette, radius 8° centered on fixation) to increase object recognition difficulty (Figure 2-1). We used a rapid event-related design where each image was presented for 1.25 s followed by a random delay between 1 and 4 s. Subjects maintained fixation while performing a 2AFC luminance decrement discrimination task on the fixation cross [91] whose timing was randomly out of sync with stimulus presentation.

2.3.3 MRI methods

Data were collected at RIKEN Brain Science Institute with a Varian Unity Inova 4T whole-body MRI scanner using a head gradient system (Agilent). We collected a T1-weighted anatomical scan (MPRAGE; TR, 13 ms; TI, 500 ms; TE, 7 ms; flip angle, 11°; voxel size, 1×1×1 mm; matrix, 256×256×180) and a T2-weighted anatomical images (TR 13 ms, TE 7 ms, flip angle 11°, matrix 256×256×180; 1 mm isotropic voxels) for each subject. We divided the T1 and T2-weighted images to correct for contrast inhomogeneities [92] and segmented this reference anatomical to generate cortical surfaces using Freesurfer [93].

We collected functional scans at 3×3×3 mm (matrix size, 64×64×27) using echo-planar imaging. Scans were collected with a TR of 1.25 s, a TE of 25 ms, flip angle 30° using sensitivity encoding (acceleration factor of 2) [94]. We showed 210 distinct images each session (105 stimuli per run, alternating between two run types). In each functional session, we collected an anatomical scan for cross-session alignment to each subject's high-resolution anatomical.

Subject 1 (S1) participated in 14 functional sessions and was shown 2539 images.

Subject 2 participated in 9 functional sessions and shown 1785 images–a subset of those shown to S1. Our analyses used the 1785 images shown to both subjects.

2.3.4 MRI data pre-processing

We recorded physiological data to reduce noise artifacts. Respiration measurements from a pressure sensor and pulse oximeter data were used for retrospective estimation and correction in k space [95]. tSENSE [96] acceleration artifacts were removed with notch filtering using mrTools. No slice time correction or spatial smoothing was performed. We corrected head motion using standard approaches [97].

2.3.5 Visual area definitions

We collected one retinotopic session for each subject [91, 98]. The imaging parameters were the same as our functional sessions (exceptions: $r = 2$, tSENSE acceleration, effective TR 1.02 s, 35 axial slices). We positioned slices perpendicular to the calcarine sulcus. Preferred angle and eccentricity for each voxel were estimated using a Fourier-based analysis and projected on the gray matter surface.

We used 6 runs for our retintopic area definitions. Two runs of both clockwise and counter-clockwise wedges were used and one run each of expanding and contracting rings. In each run we collected 168 volumes (24 volumes per cycle, 10.5 cycles). We discarded the first half cycle to minimize visual adaptation effects. While maintaining fixation, subjects performed a staircased two-alternative forced choice contrast discrimination task at fixation to maintain alertness.

Similar preprocessing was performed on the retinotopic sessions as the main experiment. After preprocessing, we time reversed (2 volume offset to correct for hemodynamic lag and improve SNR) the counter-clockwise runs and averaged together these runs with the clockwise runs. This left us with an average time-series for the ring and wedge runs. We determined the preferred angle and eccentricity phase for each voxel using a Fourier-based correlation analysis. We projected these values on the flattened gray matter surface and defined border definitions using published procedures [59].

V1, V2, V3, V3A, hV4, LO1, and LO2 were defined in the ventral stream [99, 100, 60].

2.3.6 Category area definitions

Imaging parameters for the category localizer session were the same as functional sessions (exceptions: $r = 4$, tSENSE acceleration, effective TR 1.08 s). We showed natural images matched to have identical magnitude in Fourier space to reduce differences between object categories [101]. Scrambled and intact images were shown at 14° height and width. The session was block designed with 12.9 s blocks, 13 images per block 0.75 s on, 0.25 s off.

Pre-processing for the localizer session was similar to that of our main experiment; however we applied spatial Gaussian smoothing (6 mm full width at half maximum). We created a design matrix with predictors for each of the block types by convolution with a canonical hemodynamic response function (difference of gamma functions, $x = 6$, $y = 16$, $z = 6$, where x and y were the shape parameters of the positive and negative functions and z was the ratio of the scaling parameter of positive to negative gamma functions). Using the design matrix, we fit a GLM to each subject's data individually. Using the fitted responses, we calculated a contrast for intact stimuli (scenes, faces, and natural objects) to scrambled. We defined and masked LOC using a statistical threshold of $p \leq 0.0001$ (uncorrected) and removed all voxels within the retinotpically defined areas (V1-V4). We defined PPA, OFA, and FFA using similar procedures. OFA and FFA were defined using a faces to objects contrast [102]. PPA was defined using a scenes over objects contrast [103]. Some of LOC overlapped with LO2, however it should be mentioned LOC is not a superset of LO1 plus LO2, as they are defined using entirely separate criteria (category localization versus retinotopy) [104].

2.3.7 Image responses

We used GLMs with PCA components of non-visually driven voxels as noise regressors to estimate image responses of each voxel with GLMdenoise using the package's default HRF [105], which produced for each voxel one response (GLM coefficient) for all presentations (median of 6) of each image. We computed reliability by randomly splitting the scans into two groups and estimating responses for each group. The correlation between the vectors was our estimate of split-half reliability. We discarded voxels with $r \leq 0$ (similar to [106]) and pooled voxels across subjects resulting in 536 voxels for V1, 407 for V2, 510 for V3, 379 for hV4, 123 for PPA, 192 for OFA, 292 for FFA, 234 for LO1, 299 for LO2, 111 for TOS, and 535 for LOC. Our analyses are based on the assumption that ventral visual representations are similar across subjects, based on prior work which has shown remarkable representational similarity not only across subjects but across species [56, 57].

2.3.8 Convolutional neural network architecture

We used a convolutional neural network (CNN) inspired by [6]. Our model consisted of two branches of three main layers. Each main layer contained one or more convolutional stages followed by normalization and pooling. Figure 2-2 illustrates the architecture of our network. Normalization and pooling followed the first, second, and fifth convolutional stages. We used the publicly available cuda-convnet package with minor custom modifications to train and evaluate our model [6]. Our main analyses focus principally on the outputs of the three main layers.

Each of the 5 convolutional stage was constructed using rectified linear units. Rectified linear units are a simple non-linearity of the form $f(x) = max(0, x)$ and were chosen by [6] in part because training networks with this form of non-linearity is quicker than other non-linearities. The five convolutional stages contained filters of spatial sizes 11×11, 5×5, 3×3, 3×3, and 3×3 px. Each convolutional stage had 48, 128, 192, 192, and 192 filters respectively.

We used 3 identical response normalization stages as [6]. For a given unit ${a^i}_{x,y}$ representing the activation of channel i at spatial position x, y, the normalized output

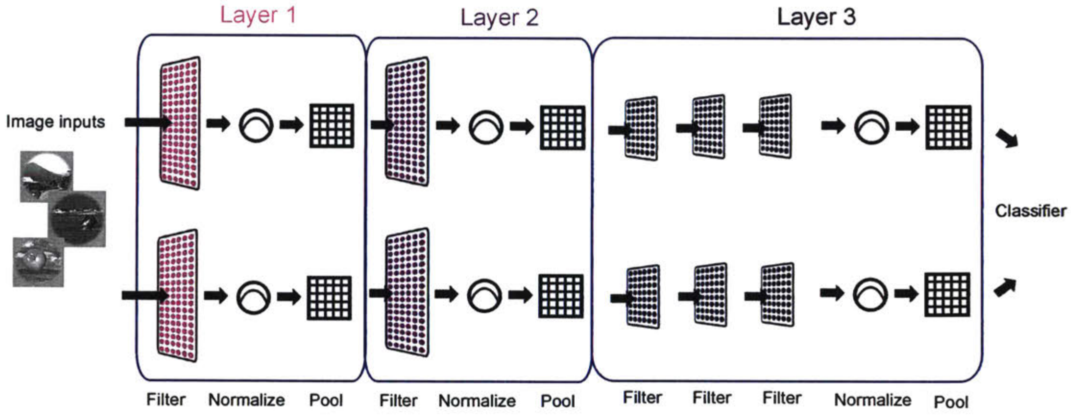

Figure 2-2: Architecture of our convolutional neural network
Beginning with the first layer and extending through until the fully connected (classifier) layer, the model contains two branches. The first convolutional layers for both branches each contain 48 filters, followed by 128 filters in the second convolutional layer, 192, 192, and 192 filters for the third, fourth, and fifth convolutional layers. We used the same filter sizes (11×11, 5×5, 3×3, 3×3, and 3×3 px) for convolutional layers 1-5 and striding parameters as [6].

is defined as,

$$r^i_{x,y} = \frac{a^i_{x,y}}{\left(k + \alpha(\sum_{j=max(0,i-n/2)}^{min(N-1,i+n/2)} \left(a^j_{x,y}\right)^2 \right)^{\beta}}$$

where n is the number of channels in the same spatial location to normalize across, and N is the number of channels in the layer. Because we initialize all convolutional weights randomly, the ordering of the channels is initially arbitrary. Like [6], we set $k = 2$, $n = 5$, $\alpha = 10^{-4}$, $\beta = 0.75$.

Our 3 max pooling stages were also defined as in [6]. Max pooling takes the maximum value across space in each channel. We used max pool windows of size 3×3 with a spatial distance of 2 units between each pooling window. Using a smaller distance between windows than the size of the windows results in overlapping pooling, which [6] observed results in a modest boost in model performance than non-overlapping pooling.

We simplified the architecture described by [6] based on a preliminary analysis of which aspects of the model influenced performance on the 2013 ImageNet challenge-set. Namely, we removed two of the middle fully connected layers (compromising the majority of the model's free parameters). Because the only remaining fully connected layer in our model was the top-layer (the classifier outputs), we did not utilize drop-out, unlike [6]. We additionally reduced the overall size of the network by reducing the input image size from 224×224 px to 120×120 px. With the training/test split of the 2013 ImageNet challenge-set we observed no significant changes in model performance after making these changes.

2.3.9 Convolutional neural network optimization

The fitting procedure used here follows that of [6]. We learned filters and bias terms for each convolutional stage and the final fully connected layer with stochastic gradient descent. Batch sizes of 128 images were used from the 2013 ImageNet challengeset. The model was not trained on any synthetic images. All normalization and pooling parameters were held fixed and chosen to match [6]. In total, 9,019,111 parameters were learned. The majority of these parameters (6,912×999 = 6,905,088) were weights for the fully connected layer, which can essentially be thought of as classifier weights for the ImageNet challengeset–the output of the fully connected layer (a vector of 999 elements) is directly normalized to give the probability that a given image belongs to each of the 999 categories. Excluding the fully connected layer, which we did not use in subsequent analyses, the remaining five convolutional stages contributed 3×11×11×48, 48×5×5×128, 128×3×3×192, 192×3×3×192, and 192×3×3×192 weighting parameters respectively per branch, in addition to 48, 128, 192, 192, and 192 bias parameters per branch, for a total of 2,113,024 parameters.

Backpropagation training was performed for several days on a single NVIDIA Titan GPU for 74 epochs. To prevent overfitting, we augmented the training set by randomly cropping 120×120 px image patches from re-scaled 130×130 px images of the 2013 ImageNet challenge-set. Weights were initiated from a zero-mean Gaussian distribution with a standard deviation of 0.01. We manually reduced the learning rate of the procedure an order of magnitude when we observed the log-probability on the testing-set no longer decreased. Three such reductions in learning rate were performed. We terminated the fitting procedure upon observing further reductions in learning rate did not produce any additional decrements in the log-probability. The final performance value of the model reached that of ~70% correct (chance = 0.1% correct) and was within error of [6].

2.3.10 Control models

We included three controls: V1-like [36], V2-like [107], and HMAX [108] models. V1-like consisted of Gabor filters at multiple scales, orientations, phases, and frequencies. V2-like consisted of non-linear conjunctions of Gabor outputs. HMAX contained hierarchical operations inspired by V1. We included an animate-inanimate RDM, created on the categorical animacy of each stimulus. The animate-inanimate RDM represents something of an upper bound to which increased categorization performance can lead to increased representational similarity for higher visual areas.

The HMAX model was built on similar principles to our CNN. It contained linear-non-linear layers involving filtering and max poolings. The architecture and training procedure of HMAX and our CNN, however differ. HMAX, for instance, contains approximately an order of magnitude less trainable parameters (10^5 vs 10^6) and is a shallower architecture. In addition, its training procedure is not gradient-based, making it somewhat less optimal in any given training regime. These properties make HMAX a reasonable intermediate control between our V1-like control model and our CNN. To give the HMAX model the best possible chance to perform, we pre-trained the model using the stimulus images used to evaluate the model (for which we have BOLD data). This is in contrast to our CNN which was never trained on any images shown to our human subjects (or even any synthetic, 3D generated images).

2.3.11 Representational dissimilarity matrices

We computed representational dissimilarity matrices (RDMs), like [56], consisting of one minus the pair-wise correlation of feature vectors (where features were GLM coefficients for each voxel in the case of brain areas and model unit outputs in the case of the model). Diagonal entries were set to 0.

Compared to other studies [56, 57], we used a far larger stimulus set where each object appears in multiple images shown in different positions, orientations, and scales. Because we were interested in the emergence of object perception, we created RDMs of object-averaged response vectors where we average features across images representing the same object. The object-averaged RDMs were also necessary to increase the amount of signal in our data — our stimulus set was purposely designed to be very difficult for observers to recognize the objects in order to expose the key computational aspects of invariant object recognition. Even when given infinite viewing time, there are many images in our stimulus set that human observers cannot recognize due to extreme variations in pose, orientation, and scale.

Because responses in each imaging voxel likely result from the activity of multiple neurons with different feature selectivities, we used a linear re-mapping of model features (c.f. [57]). We computed the correlation between model layers RDMs and visual response RDMs using a linear re-mapping of model features to match a given visual area's RDM–each model layer and visual area pairing had its own set of weightings. The advantage of this approach is that it does not require model features be precisely synonymous with voxels which reflect large collections of neurons with potentially varying selectivities. The disadvantage of re-mapping is that it may be prone to overfitting, which we address with cross-validated bootstrapping and regularization. To estimate effect sizes, we used cross-validated bootstrapping which has the advantage of estimating our fitting reliability but is disadvantageous in that it requires us to fit on random subsets of the dataset rather than all of it. Each training set consisted of 1000 randomly selected model outputs to 15 images for each of 64 objects (960 images total). Model outputs for the remaining 12 images per object were used for testing. We found the vector w that maximizes $corr(RDM(V), RDM(X \circ W))$, where $corr()$ is the Pearson correlation, $RDM()$ is the vector of pair-wise row correlations, V is the matrix of object-averaged voxel responses (objects by voxels), X is the matrix of object-averaged model features (objects by 1000), W (objects by 1000) consists of rows of w, and $\circ$ represents point-wise multiplication. We find w using the L-BFGS-B algorithm [109] for 1 iteration (to both reduce computational time and as a form of early stopping to prevent overfitting). We report the average correlation on the testing set over 100 bootstraps (Figure 2-3) and 10 bootstraps (Figure 2-5). We used this procedure to calculate correlation values for all model layers as well as for all control models. Our linear re-weighting procedure is closely related but not identical to [57]. [57] fit one weight per layer or model instead of per feature. With the correct normalization, squared Euclidean distances are proportional to correlation distances and non-negative least-squares on this quantity should maximize the RDM correlation distance like the method we used here.

Our linear re-mapping procedure did not overfit to the training images (train and test correlations were similar), due to regularization (early stopping). We additionally verified that if we were to overfit the linear re-mappings on the training images, that this would not automatically result in high correlations on the testing images. If this were the case, the RDM correlations would effectively measure how invariant the model representation is. Empirically, we found that overfitting, for instance, pooling layer 3 to training images from LOC produced lower correlations on LOC testing images, indicating that our procedure would not merely quantify the invariance of model layers, even if it were substantially overfitting the weightings to the training set.

We were not able to reliably calculate split-half explainable variance estimates for this linear re-mapping procedure due to the difficulty of fitting weights on smaller fractions of our data. However, these estimates were not critical to the hypotheses tested in this study because we were comparing the relative ranking of model predictivity for each visual area (ex. layer X explains visual area A significantly more than layer Y). To avoid the problem of finding linear re-weightings using smaller sub-sets of our data, we instead computed noise ceilings and percent explained variance values (Figure 2-6) without using the weighting procedure described above. Noise ceilings for each visual area were computed by splitting the runs of our data into two non-overlapping groups. With each group, we estimated stimulus responses (beta weights) using the procedure described above (see the Image responses section) and computed object-averaged RDMs for each visual area. We used the correlation between the RDM from each of the two groups as our noise ceiling for percent explained variance estimates (Figure 2-6).

2.3.12 RDM statistical analysis

Using the bootstrapping above, we computed p-values testing if Layer A better explained visual area X's RDM than Layer B (where $A = 1$ and $B = 3$, $X =$V4, for example). We use the notation $p_{LA<LB}$ to denote the p-value of $r_{V,A} < r_{V,B}$, where $r_{V,A}$ is the testing-set Spearman correlation of layer A and visual area V's RDMs averaged over bootstraps. We use Fisher's r-to-z transformation using [110]'s approach to compute p-values for difference in correlation values [111]. The approach tests for equality of two correlation values from the same sample where one variable is held in common between the two coefficients (in our case, an RDM of a given visual area). We report p-values which are not corrected for multiple comparisons. Our approach bootstraps over independent stimulus samples and avoids problems that can arise from randomly sampling RDM matrices directly. Direct sampling of the RDM (ex. randomly sampling elements from it) can be problematic because two such random samples are not independent–a single stimulus contributes to multiple elements in the RDM matrix [112].

Spearman rank correlations are known to be biased for RDMs containing many tied ranks and can produce artificially high correlations [112]. While the animate-inanimate control RDM has many tied rankings, none of our model or visual area RDMs contain tied ranks. For this reason, using Spearman correlations with the animate-inanimate RDM may produce misleadingly high correlations, particularly for higher visual areas. As noted in [112], Kendall's Tau correlation penalizes tied ranks, however, empirically, for our data-set it does not produce qualitatively different results. That is, even with Kendall's Tau correlation, the animate-inanimate RDM significantly out-performs all model layers (ex. for a single bootstrap we observe a Tau value of 0.328 for animate-inanimate to LOC vs. a Tau value of 0.121 for layer

3 to LOC).

2.3.13 Classification

We assessed model and neural recognition performance with cross-validated linear support vector machines (SVMs). Classifiers were trained on stimulus category of individual image responses. Training consisted of 20 random presentations of each object and testing consisted of the remaining presentations. We report median accuracy over 20 bootstraps. We set the classifier regularization "C" parameter equal to 0.0005 and computed significance by a one-tailed Welch's t-test. We have not performed corrections for multiple comparisons.

2.3.14 Performance vs. fitting

During ImageNet optimization, we measured model and neural similarities. At 100 gradient updates (checkpoints) spaced evenly through optimization, we computed RDM correlations using the procedure above. We sampled 100 points spaced evenly over the range of model performance values and plotted the average correlation over model checkpoints within 0.10 accuracy of each sampled point.

2.4 Results

We optimized a convolutional neural network model for object recognition on a challenging image-set [113] to test the extent it matched the human visual system. After optimizing using backpropagation, the model achieved ~70% accuracy (chance = 0.1%) on ImageNet, and comparable although slightly reduced performance relative to humans, consistent with previous work [6, 10].

Emergence of categorical information was evident in model and human representations. We computed object-averaged RDMs [56] for visual areas and model layers (see Materials and Methods). Each entry in an RDM is a measure of how dissimilarly a pair of objects are represented. Arranging the stimuli by category, we observe the

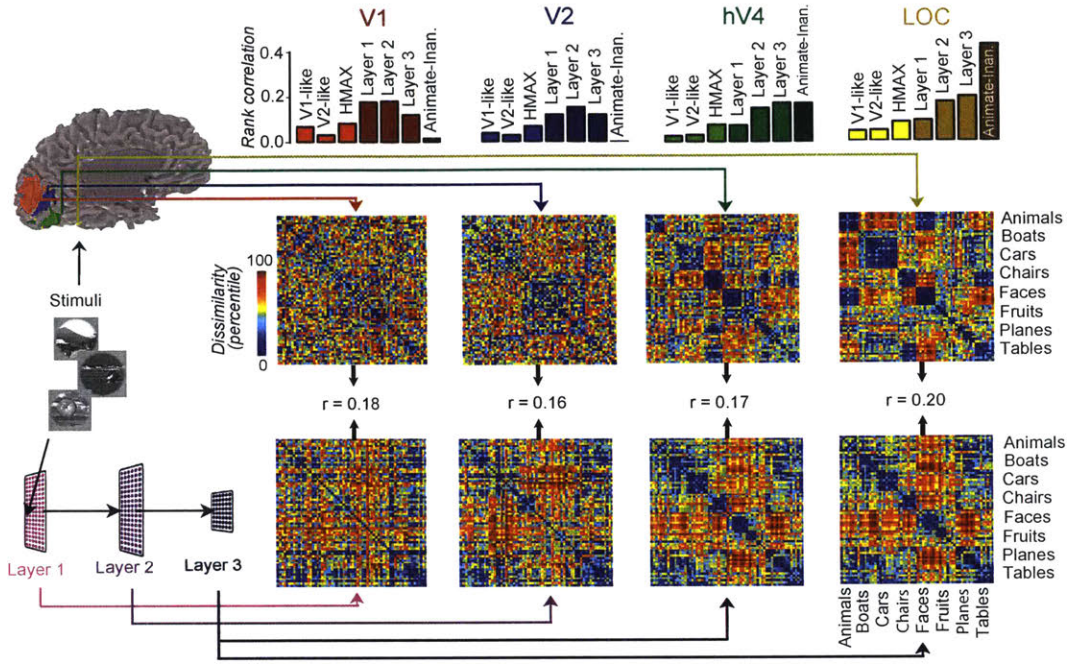

Figure 2-3: Model and neural representations

Left, Human functional imaging data and model responses to the same stimuli were used to compute RDMs at different levels of the visual system (top row) or layers of the model (bottom row). Increasing block-diagonality of the RDMs from V1 to LOC and from Layer 1 to Layer 3 illustrate emergence of categorical representations. Rank correlations between model layers and visual area RDMs **Top,** showed better correspondence than control models (V1-like, V2-like, and HMAX). Bars indicate SEM over bootstraps (see Materials and Methods).

emergence of block-diagonality (Figure 2-3). Blocks correspond to the emergence of categorical tolerance through the ventral stream, as within-category similarities are increasingly abstracted despite the high levels of variation in the stimuli. The RDMs of the model (Figure 2-3) also evidence emergence of categorical information.

We quantified recognition performance in model and visual areas by training support vector machines (SVM) to decode the category of each stimulus response (Figure 2-4). We observe increasing performance as we move from lower to higher model layers ($p_{L2>L1} = 6.3 \times 10^{-48}$; $p_{L3>L2} = 3.5 \times 10^{-38}$; see Materials and Methods: Classification) and increased performance as we move from posterior to anterior areas (as shown in Figure 2-4; $p_{V2>V1} = 0.0065$; $p_{hV4>V2} = 5.3 \times 10^{-57}$; $p_{LOC>hV4} = 2.4 \times$

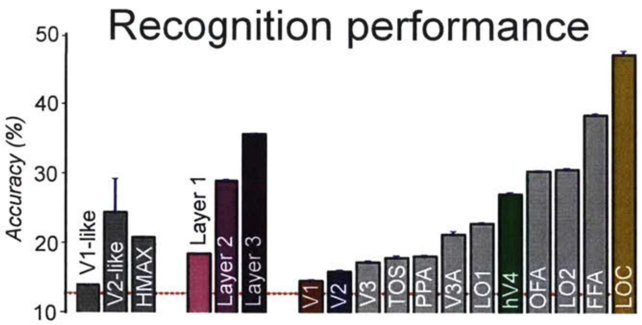

Figure 2-4: Object recognition performance
Accuracy for each model and visual area was computed with a cross-validated linear support vector machine (chance = 12.5%; dashed red line). The same training/test procedure was used for model and neural responses.

10^{-71}). V1-like and HMAX models generally perform worse than the layers of our model ($p_{L3>HMAX} = 4.1 \times 10^{-63}$; $p_{L2>HMAX} = 3.2 \times 10^{-37}$). V1-like performs similarly to the fMRI V1 responses (but worse than all of our model layers–$p_{L3>V1-like} = 8.4 \times 10^{-78}$; $p_{L2>V1-like} = 1.2 \times 10^{-63}$; $p_{L1>V1-like} = 5.9 \times 10^{-29}$). HMAX performs in between V2 and hV4 responses ($p_{HMAX>V2} = 1.1 \times 10^{-27}$; $p_{hV4>HMAX} = 4.4 \times 10^{-38}$).

We found correspondence between model pooling layers and visual areas (see Tables 2.1 and 2.2). Early areas were best explained by early layers and later areas by later layers (Figures 2-3 and 2-6, e.g. compare layer correlations of V1 to LOC). V1 was best explained by lower–layers ($p_{L1>L3}$= 0.0058; $p_{L2>L3}$= 8.9×10^{-4}; see Materials and Methods: RDM statistical analysis), and LOC was best explained by higher layers ($p_{L3>L1}$= 5.4×10^{-6}; $p_{L3>L2}$= 0.13; $\mathrm{p}_{L2>L1}$= 9.9×10^{-7}). We observed intermediate visual areas, such as V2 and hV4, following this trend. V2, for instance, was better explained by the middle Layer 2 than the top layer ($p_{L2>L3}$= 0.047).

Our model exhibited higher similarity to the ventral stream than several control models: a V1-like model [36], a V2-like model [107], and HMAX [108] (ex. for V1 $p_{L1>HMAX}$= 1.3×10^{-4}; for V2 $p_{L1>HMAX}$= 0.026, for hV4 $p_{L2>HMAX}$= 1.8×10^{-3}, and for LOC $p_{L2>HMAX}$= 1.5×10^{-4}; see Materials and Methods: RDM statistical analysis). HMAX, V1-like, and V2-like models predicted hV4 and LOC RDMs

Variable	Data structure	Type of test	p-value
$p_{L2>L1}$	Normal distribution	Welch's one-tailed t-test	6.3×10^{-48}
$p_{L3>L2}$	Normal distribution	Welch's one-tailed t-test	3.5×10^{-38}
$p_{V2>V1}$	Normal distribution	Welch's one-tailed t-test	0.0065
$p_{hV4>V2}$	Normal distribution	Welch's one-tailed t-test	5.3×10^{-57}
$p_{LOC>hV4}$	Normal distribution	Welch's one-tailed t-test	2.4×10^{-71}
$p_{L3>HMAX}$	Normal distribution	Welch's one-tailed t-test	4.1×10^{-63}
$p_{L2>HMAX}$	Normal distribution	Welch's one-tailed t-test	3.2×10^{-37}
$p_{L3>V1-like}$	Normal distribution	Welch's one-tailed t-test	8.4×10^{-78}
$p_{L2>V1-like}$	Normal distribution	Welch's one-tailed t-test	1.2×10^{-63}
$p_{L1>V1-like}$	Normal distribution	Welch's one-tailed t-test	5.9×10^{-29}
$p_{HMAX>V2}$	Normal distribution	Welch's one-tailed t-test	1.1×10^{-27}
$p_{hV4>HMAX}$	Normal distribution	Welch's one-tailed t-test	4.4×10^{-38}

Table 2.1: Layer-wise t-test statistics

Variable	Data structure	Type of test	p-value
$p_{L1>L3}$	Two dependent correlations	Asymptotic z-test	0.0058
$p_{L2>L3}$	Two dependent correlations	Asymptotic z-test	8.9×10^{-4}
$p_{L3>L1}$	Two dependent correlations	Asymptotic z-test	5.4×10^{-6}
$p_{L3>L2}$	Two dependent correlations	Asymptotic z-test	0.13
$p_{L2>L1}$	Two dependent correlations	Asymptotic z-test	9.9×10^{-7}
$p_{L2>L3}$	Two dependent correlations	Asymptotic z-test	0.047
$p_{L1>HMAX}$	Two dependent correlations	Asymptotic z-test	1.3×10^{-4}
$p_{L1>HMAX}$	Two dependent correlations	Asymptotic z-test	0.026
$p_{L2>HMAX}$	Two dependent correlations	Asymptotic z-test	1.8×10^{-3}
$p_{L2>HMAX}$	Two dependent correlations	Asymptotic z-test	1.5×10^{-4}

Table 2.2: Layer-wise z-test statistics

approximately as well as Layer 1 of our model. For earlier visual areas, the control models were significantly worse at predicting the neural RDMs than any layer of our model (see aforementioned statistics). Our model exhibited lower correlations than the animate-inanimate RDM in LOC. However, unlike other controls, the animate-inanimate RDM does not represent the outputs of an image-computable model. The animate-inanimate RDM represents something of an upper bound in terms of how far we might expect increased performance optimization to lead to increased neural fitting of higher visual areas. It should be noted that we have not arranged the rows and columns of our RDMs in a way that visually highlights the animate-inanimate distinction observed previously [56]. However, the animate-inanimate RDM correlations are a quantitative measure of this phenomenon and the high correlations of higher visual areas (ex. LOC) to this matrix indicates consistency with previously reported findings [56].

If recognition performance is key to driving correspondence between model and brain representations, then improving model recognition performance should also improve correlations between model layers and visual areas. We found that the model's correlations increased as a function of its optimization on ImageNet (Figure 2-5). For each step the model took toward better performance, it also became increasingly similar to neural data. As is known from previous work [88, 114], spatial receptive fields (pooling of inputs) plays a significant role in voxel responses of early vision. We also observe this — Layers 1 and 2 have higher RDM correlations with V1 than Layer 3 even before the model has been highly optimized. However, the pooling structure of our model alone cannot explain these results since as the model becomes optimized, its similarity to V1 and other areas increases, despite the pooling of the model remaining fixed. LOC is not best explained by Layer 3 until the model has been well-optimized–that is, optimization drives Layer 3 above Layers 1 and 2.

We additionally analyzed intermediate convolutional and normalization stages (Figure 2-6) by computing their object-averaged RDM correlations to each of the visual areas. We observed that the intermediate convolutional and normalization stages roughly fall between the pooling layers in terms of their mapping to each

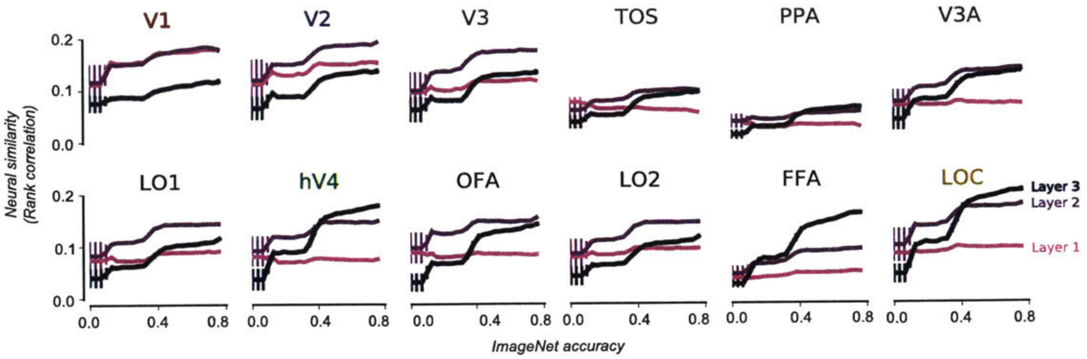

Figure 2-5: Optimization at object recognition performance vs. predictivity Correlations between RDMs of each model layer (different colors; light purple: Layer 1, medium: Layer 2, dark: Layer 3) and visual area (different graphs) are shown as a function of model performance on ImageNet taken at different "checkpoints". A positive trend indicates that, as the model becomes optimized on ImageNet recognition, it is better able to explain neural responses. Vertical bars indicate SEM over checkpoints (they become eclipsed by the width of the line plot on the far right of the plots).

visual area. For practical reasons, Figure 2-6 presents the unweighted RDM correlations. Empirically, we observed that randomly selecting 1000 features is insufficient to produce stable RDMs from these model stages. Therefore, we present the unweighted RDM correlations using all of the feature dimensions for each layer because computing many more than 1000 feature weightings was infeasible. This change was necessary because the convolutional and normalization stages contain four to nearly ten times more feature dimensions than the pooling layers. Because we did not utilize feature re-weighting, we were able to reliably estimate noise ceilings for these correlations. Determining noise ceiling for correlations where we used feature re-weighting (Figures 2-3 and 2-5) was infeasible because it requires estimating the weights on smaller subsets of the data for which we were unable to learn stable weightings.

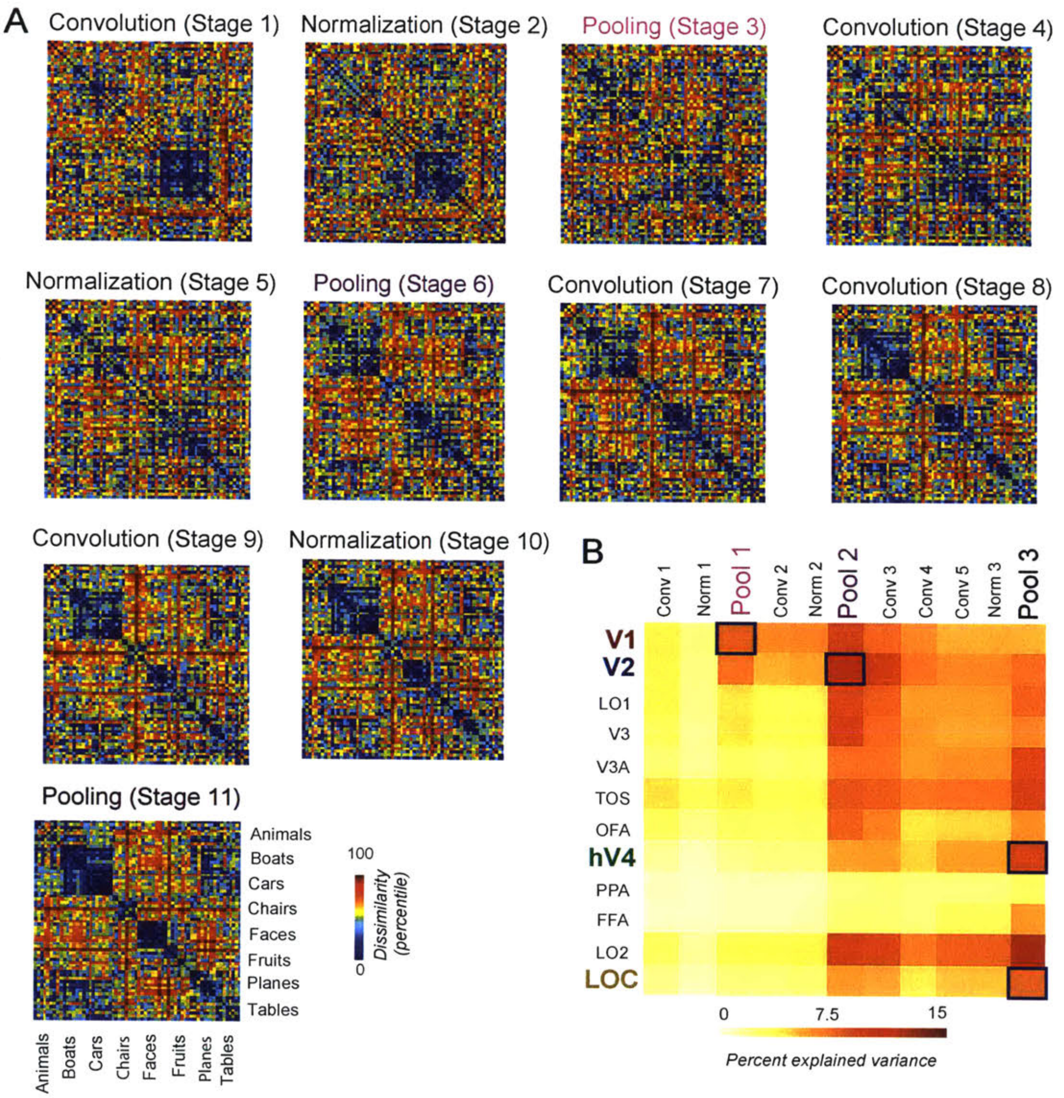

Figure 2-6: Object-averaged RDMs for all model stages

A, Shown are the unweighted model RDMs–they are not re-weighted to any of the fMRI visual area responses, unlike Figure 2. The pooling stages represent pooling layers 1 through 3 which were used in our main analyses. **B,** Shown are the percent explained variances between each model stage and visual area. We did not re-weight the model features in this analysis–instead of sampling 1000 random features we used all features from a given model stage. Blue boxes indicate, for each selected ROI, which model layer exhibited the highest correlation to the ROI.

2.5 Discussion

By analyzing human BOLD responses to hundreds of images, we were able to compare representations of our deep convolutional neural network to those of early, intermediate, and late visual areas simultaneously, thus extending previous work [10, 115] both to humans and to the hierarchy of topographically and functionally localized visual areas (c.f. [57, 58]). We found that a deep convolutional neural network optimized for object recognition had representational similarity to human ventral stream visual areas in a hierarchically consistent fashion — early layers best predicted early visual areas and later layers best predicted later areas. The intermediate convolutional and normalization layers residing between the pooling layers exhibited similar, but not as precisely ordered mapping (ex. the second convolution and normalization layers produce very similar RDMs; Figure 2-6B). The hierarchical correspondence between the network pooling layers and human cortical visual areas increased as the model's recognition performance was optimized to perform object recognition, suggesting that the functional constraint of object recognition performance was a key component for representations to emerge that resemble ventral visual stream representations. Taken together, our results suggest that biologically plausible computations (convolution, threshold non-linearities, pooling and normalization, [72]) coupled with the top-level constraint of image recognition performance is sufficient to produce hierarchical representations similar to those found in the human visual cortex.

Our analysis of visual representations averaged BOLD responses and model representations to the same object shown from different views, thus stressing object properties common to different viewpoints over ones that are different between viewpoints. Examining responses to individual exemplar images with a single viewpoint [57] might give insight into the development of tolerant representations, however, our stimulus set did not include enough repeats of the same image to allow for split-half reliability sufficient to analyze without averaging across all views of an object. A potential concern with our object-averaging procedure is that it might artificially favor stronger representational correspondence between the model and more view tolerant cortical areas [116, 5, 117]. However, we did not find this to be the case. Instead, correlations were of comparable magnitude across V1 to LOC to the model, what differed was which layer best correlated with each area. We note that correspondence after object-averaging does not necessarily mean that all visual areas or model layers have highly tolerant representations; incidental properties of objects that are still not averaged out across different views might also drive correlations between the model and cortical responses.

The notion that the visual system is hierarchically organized [118] suggests that intermediate visual areas like V2 and V3 contain intermediate representations, but intermediate in what sense? Our results demonstrate that similar intermediate representations naturally emerge from a deep convolutional network as object recognition performance is optimized, suggesting that the top-level object recognition constraint is sufficient to constrain these intermediate representations. An alternative is that similarity to intermediate areas might only emerge when each model layers are independently optimized for a relevant task (e.g. edge detection for the first model layer, curvature conjunctions for middle layers, object recognition for the higher layers). Nothing in the training of the neural network forced representations to conform to the intermediate representations in visual cortex - the neural network could have learned to generate categorical representations through completely different intermediate mechanisms than those in visual cortex, but our evidence suggests otherwise. While our approach differs from others who have sought to understand what explicit computations might be done in intermediate areas, such as for curvature [119, 120, 3], angles [121] or for conjunctions of orientations [71, 122, 123, 124] or other features [125, 126], we note that our results do not preclude such an understanding of intermediate visual areas. To what extent the intermediate layers of the model can be characterized as making such explicit computations is a matter of continued investigation which could, in principle, be done by analysis of the receptive field properties of neural network units [127].

We found that training on object recognition performance was sufficient to drive representational similarity between the model and the human visual system suggesting that model performance and not the specific model architecture was the important factor. Indeed previous analyses [10] showed the strongest correlations of model to monkey physiology data was driven by object recognition performance rather than any specific model parameter. This suggests that the exact formulation of the neurally inspired operations [128, 1] like convolution (linear RFs), rectification (spike threshold), normalization and pooling in a layered architecture are less important than the top-level object recognition constraint. Therefore, if other hierarchical models of visual processing similar in architecture to ours, such as HMAX [75], could be trained to have much higher recognition performance, its representations might become more predictive of the hierarchy of the human ventral stream.

While we trained our network solely for object recognition, human visual areas including in the ventral stream likely subserve a multitude of visual functions, and moreover some make connections into dorsal stream areas in parietal cortex thought to subserve other functions such as action planning [62]. Why then is object recognition performance sufficient to create representations in our model similar to the visual cortex? Object recognition performance may instantiate representations that also support read-outs for other object properties such as position or 3D orientation that might be important for visual functions such as action planning. Alternatively, but not mutually-exclusively, training networks to perform multiple different tasks may better constrain representations to match across multiple visual areas in humans.

There are many facets of visual representation in human visual cortex which are not adequately predicted by the model. For instance, we found that the animate-inanimate distinction was a better predictor of higher visual area responses than our model. However, animacy, like many high-level semantic categories [102, 129, 103, 130] is not (yet) image-computable and therefore does not represent a model of visual processing. It may be that top-down input representing linguistic, semantic and other cognitive factors or high-level conjunctions and associations between complex stimuli may be required to fully explain high-level representations, particularly for complex representations in the most anterior parts of the ventral stream [131]. It may also be the case that additional task constraints such as better model training performed on even more realistic object-recognition challenges than ImageNet categorization [10] are needed to improve the model correspondence to human visual areas. Our model also does not yet predict the discrete changes in representation for successive and neighboring areas in human visual cortex for low-level visual features like decrements in image contrast [132] or motion coherence [133]. Nor does it predict spatially compact clusters of similar representations such as those found in face patches [102, 81]. Human object and, in particular, face recognition display particular phenomenology [82, 134, 135] that may be different from the phenomenology and the types of errors that are made by deep convolutional networks. Thus suggesting that deep convolutional networks using current training regimes are not recapitulating all aspects of human object vision and representation [136, 137]. Nonetheless, our results here suggests that starting with biologically inspired computations and a top-level description of just one important function of the visual system can provide a sufficient starting point for explaining representations in the whole set of hierarchical visual areas we examined in human cortex. Our results thus challenge the idea that each visual area in the hierarchy of visual areas should be understood as having a cicumscribed and easily definable function.

Chapter 3

Statistical properties of high-performing CNNs and their relation to natural image statistics and unsupervised learning

Recent advances in computer vision and machine learning have made it possible to train large neural network models to accurately recognize objects in natural images from thousands of categories. To achieve high performance, these networks are typically trained on millions of human-labeled images. Here we analyze several statistical properties of filters from high-performing supervised models. We find that statistics of the learned filters exhibit strong similarity to the output statistics of the preceding layer, starting with the natural image statistics at the first layer[1]. By constructing models optimized to match these natural statistics, we then test the sufficiency of these statistics for producing recognition performance on a challenging recognition task. We find that we are able to maintain the performance of the fully supervised model at the first two layers of high-performing networks. With three layers and beyond, these same statistics also maintain high performance, but no longer match that

[1] Work done in collaboration with Daniel Yamins and James DiCarlo

of the original supervised model. The work presented here provides a step towards better unsupervised learning procedures.

3.1 Introduction

Current state-of-the-art computer vision algorithms rival the ability of humans to recognize thousands of object categories in natural image photographs [127]. These models consist of large convolutional neural networks containing tens of millions of parameters and are usually trained by gradient-based backpropagation procedures that require millions of human-labeled images to prevent overfitting [138, 6]. For tasks where limited labeled data exists (ex. medical diagnostics, sensors outside the visible spectrum, etc.), training these networks can be difficult. Moreover, while these large supervised networks have been shown to predict the fully-developed adult neural representations in brain areas that are responsible for visual object recognition [10, 11], it is very unlikely that the heavily supervised training algorithms used to create these networks accurately describe the neural learning rules by which brains build these representations during development. Thus, finding unsupervised or semisupervised learning procedures that equal the performance of supervised procedures would be of great relevance, as such algorithms could be flexibly and efficiently deployed in a wide variety of uncontrolled real-world tasks, and potentially yield a better understanding of the biological visual learning.

A variety of unsupervised learning rules have been put forth in the literature, many of which are capable of producing high-performing (but not currently state-of-the-art) models on challenging recognition tasks. These learning rules often take the form of simple mathematical formulations that optimize filter parameters to produce one or another pre-determined statistical property, (ex. reconstruction/auto-encoding [139, 140, 141], slowness [22], sparsity [45], denoising [142], etc). Historically, however, these ideas were proposed before much of the recent capacity of supervised convolutional neural networks was fully realized. With the advent of high-performing convolutional neural networks, we take a different approach. We ask: what statistical properties of the supervised filters are necessary and sufficient to replicate the performance of already high-performing models?

We describe an initial set of summary statistics of convolutional filters, based on the observation that second-order correlation statistics describing each layer's filters closely resemble the correlations in the input statistics to that layer (or natural image statistics in the case of the first layer). We test the extent to which filters optimized to match these key summary statistics produce high-performing models. We find that for the first two layers, these statistics are a sufficient characterization for producing models with comparable performance to supervised models. Above the second layer, we find that a performance gap opens up, suggesting that it will be necessary to identify additional statistical constraints to produce high-performing models.

3.2 Prior work

Most previous work on unsupervised feature learning is based around the general idea of optimizing for network outputs with specific properties (ex. slowly varying features [22], features with denoising or reconstructive properties [142, 139, 141], sparseness, etc. [143, 144]). These objective functions were often first derived from intuitions about what constraints might have shaped the primate visual system (ex. sparseness from energy constraints [45]) or based on suspected properties of high-performing models (ex. autoencoding or formulations of mutual information). While many of these approaches can be used to produce high-performing models or have provided insight about primate vision, no known unsupervised procedure is currently capable of producing a model with performance near that of the current state-of-the-art supervised models. Much less work has been done however on exploring objective functions on the statistical properties of filters directly, rather than the properties of their outputs. This discrepancy is partly due historically to the fact that, until relatively recently, the full potential of supervised convolutional neural networks had not been fully realized, therefore no good target existed for which these properties could be studied. Nevertheless, many ideas have been put forth (ex. ICA with sparsity constraints [45, 145, 146], efficient coding, etc. [147]) that, when implemented, produce V1-like properties at early model layers [45], and in some cases also produce models that perform well on digit recognition and texture classification problems [148]. Additional regularities of natural image statistics (eg. spectral properties [149], spatial correlations, joint distributions of pixel intensities [147]) have been observed, but no known study has explicitly incorporated this information into a learning procedure capable of producing a deep, high-performing network.

While many intuitive and conceptually elegant descriptions have been found to characterize filters in the first layer of neural networks (ex. [45]), describing the properties of higher-level model filters has proven more difficult. In part due to this difficulty, [127, 15] has proposed visualizing feature projections from intermediate and higher-level model filters. These approaches provide insight into the salient properties and features of images these model layers are representing, but do not directly provide insight into the regularities and properties of the filters themselves. It has been suggested for some models that much of the information contained in these filters is redundant and predictable by a relatively small proportion of filter parameters [150]. Here we present a formulation of summary statistics that evidences clear structure in intermediate model layer filters—structure that is predicted by the input image statistics of each layer.

3.3 Summary statistics

We build on prior work which has sought to understand and improve deep convolutional neural networks by examining the output properties of the network. However, instead of searching for regularities in the output properties of high-performing networks (sparsity, reconstruction, denoising, etc.), here we have searched for statistical regularities in the filters themselves. The following three sub-sections describe summary statistics that, when matched, significantly increase model performance relative to untrained baseline.

3.3.1 Second order statistics

The main statistical constraint we work with are second order correlational statistics. We find that the patterns of filter second order statistics observed in backpropagation-trained models closely matches the natural image statistics of the dataset the model was trained on.

The filters of a convolutional neural network (CNN) can be thought of as a four-dimensional tensor, in which the first dimension is the *channel* dimension, representing the multiple dimensions in the input data; the second and third dimensions represent *spatial* dimensions over which the convolution operation is performed; and the last dimension is the *filter* dimension, representing the multiple axes of output that the filters define. Thus, for a CNN containing N layers, the i-th layer filters are a real-valued tensor of shape (nc_i, fs_i, fs_i, nf_i), where nc_i is the number of channels, fs_i is the filter spatial shape, and nf_i is the number of filters at layer $i \in [1, \ldots, N]$. In our case, the first layer takes input from 3-channel color images, so $nc_1 = 3$. Because the filter dimensions of one layer feed into the channels of the next layer, $nf_i = nc_{i+1}$ for all i.

To define second-order filter statistics (Fig. 3-1), we reshape F_i so that the channel and spatial dimensions are unravelled, leaving a two-dimensional tensor $\hat{F_i}$ (e.g a matrix) with shape $(nc_i \cdot fs_i^2, nf_i)$. The *second order statistics of* F_i are then defined as the pairwise Pearson correlation of the rows of this matrix, e.g.

$$M_i[j,k] = corr(\hat{F}_i[j], \hat{F}_i[k]).$$

The matrix M_i is of shape $(nc_i \cdot fs_i^2, nc_i \cdot fs_i^2)$.

Second order natural image statistics are defined in a similar manner to the statistics for the filters. Randomly sampling K image patches of size $fs_1 \times fs_1$, produces a tensor of shape $(3, fs_1, fs_1, K)$. The tensor is then reshaped into a 2-tensor just as for the filter values described above and then z-scored across the first dimension (removing luminance fluctuations across patches). The row-wise correlations are then computed to produce a matrix N_0 of shape $(3 \cdot fs_1^2, 3 \cdot fs_1^2)$.

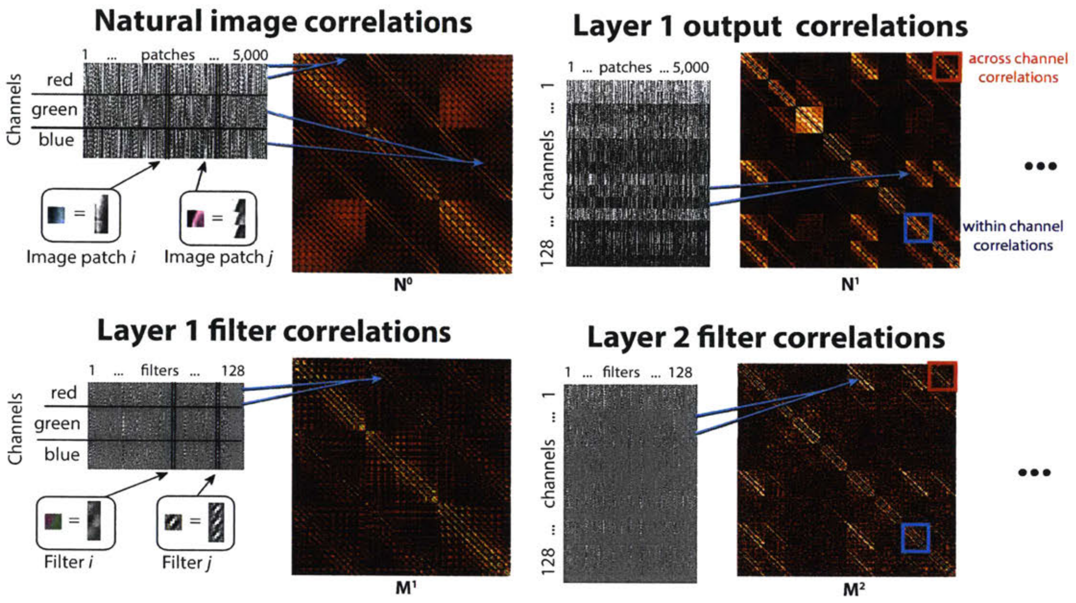

Figure 3-1: Procedure for computing filter and natural image statistics
Each convolutional layer can be represented by a four dimensional tensor: (nc_i, fs_i, fs_i, nf_i), where nc_i is the number of input channels (ex. for layer 1, this is 3 for the input channels red, green, and blue), fs_i is the size of the filter in pixels, and nf_i is the number of filters. The filters of layer $i-1$ correspond to the channels of layer i. Similar statistics can be computed for the input data that each layer receives. For example, with the first layer a four dimensional tensor can be created with image patches, with identical dimensions to the filter matrix except for the last dimension which represents the number of patches used to compute the statistics—empirically we find that any random selection of approximately 5,000 image patches produces almost exactly identical statistics. We reshape these 4-tensors into two-dimensional matrices of size $nc_i * fs_i^2 \times nf_i$ and $nc_i * fs_i^2 \times K$ and calculate second order statistics as the pair-wise row correlations. For the second layer, a cropping of the full correlation matrices are shown to illustrate detail. Correlations within a single channel are shown as blocks of size ${fs_i}^2 \times {fs_i}^2$ along the diagonal (ex. $11^2 \times 11^2$ for layer 1, $5^2 \times 5^2$ for layer 2–blue boxes in the above figure). Correlations across pairs of channels are shown as blocks of size ${fs_i}^2 \times {fs_i}^2$ along the off diagonals (ex. red boxes in the above figure). Within each of these $fs_i^2 \times fs_i^2$ blocks are blocks of size $fs_i \times fs_i$ representing the correlation of individual filter pixels with all other pixels. Diagonals in these sub-sub blocks represent the extent to which a given spatial location is correlated across two pairs of channels.

This same idea can be repeated for each layer. In CNNs, the inputs to layer i are the outputs of layer $i-1$. For K original inputs of square shape, the output of layer 1 will have shape (nf_1, o, o, K), where o is the output-size after layer 1 (which will be dictated by the original input image size, border effects, padding, and potential pooling operations). By randomly selecting spatial patches of size $fs_2 \times fs_2$ for each input, we can compute N_1 exactly as M_2 was computed. The *i-th layer second-order natural image statistics*, which we will denote, N_i, can be defined iteratively, and will be the same shape as the filter statistics M_i. Empirically, we find that we do not need K (the number of sampled image patches) to be larger than 5,000 for the values in each N_i to stabilize.

For both natural imagc and filtcr statistics, the basic structure of the correlation matrices shows significantly more correlation along super-diagonals close to the main diagonal, capturing the fact that nearby pixel pairs have more similarity than further ones (Fig. 3-2). The rate at which this correlation falls off is characteristic of the image data. However, there are channel pairs for which the correlation is high, capturing the fact that those channels share some common features relative to the input data. At the first layer, this simply expresses the fact that the red, green, and blue color channels in images are strongly related; at higher layers, this expresses a more sophisticated higher-order similarity structure.

A key "empirical" observation that we base our work on is that, for filters trained via back-propagation on complex categorization tasks, N_{i-1} bears a striking resemblance to M_i. Those channel pairs for which the natural image statistics (in N_{i-1}) show high correlation are invariably those for the filter statistics at the next layer (in M_i) show high correlation. In other words, the supervised learning process causes filters to learn some aspects of the natural statistics of the environment, as measured by these second-order statistic metric. However, the relationship between N_{i-1} and M_i is not trivial, because certain features of the natural image statistics seem to be reliably changed in the filter statistics (Fig. 3-2).

We have observed that filters from supervised models trained on different imagesets have significantly different statistics, but that these will correspond to differences

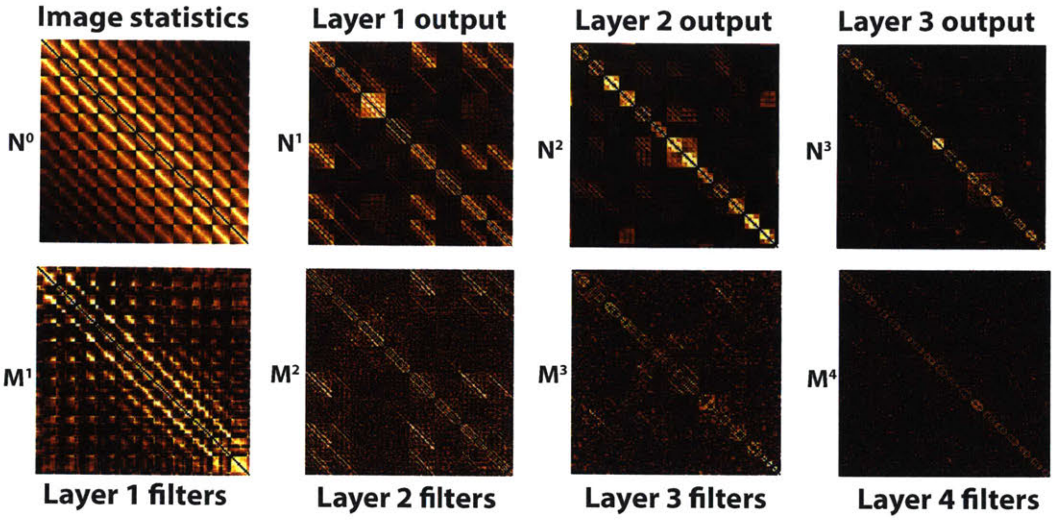

Figure 3-2: Filter and natural image statistics for each convolutional layer
The bottom row represents the magnitude of second order filter statistics of a CNN trained on ImageNet and the top row represents the magnitude of second order output statistics of the previous layer (i.e., what each corresponding filter layer sees as input). See Fig. 3-1 for details on how these statistics are computed. Here we show subportions of the correlation matrices to emphasize detail.

in the natural image statistics in the image-sets themselves. For example, filters trained on ImageNet categorization (Fig. 3-2) will correspond to ImageNet natural statistics, while filters trained CIFAR (not shown) will have second-order statistics closely reflective of the correlation statistics in the CIFAR dataset. We have also observed that the second-order filter statistics appear to be determined in a largely layer-wise fashion, meaning that for a layer i in a k-level model, the filter statistics M_i will be very similar to those from the $k-1$-level model in which the top layer has been removed. These characteristics suggest that matching the second order statistics of filters might be a fruitful route toward an unsupervised learning procedure.

Two key questions, however, are (1) *unsupervised-ness*: how does one compute M_i from N_{i-1}, since the relationship is not trivial?, and (2) *utility for performance*: even if one was able to perfectly predict M_i from the unsupervised data, is matching M_i useful for increasing the performance of a model above random baseline? In this work, we address question (2), because without performance utility additional considerations are irrelevant.

We addressed this question by optimizing filter values to match the M_i statistics of a backpropagation-trained model. We then substituted these filters back into the model and computed performance on ImageNet classification tasks (see Section 3.4.1 below). The performance of the statistically-matched models was not significantly higher than the random baseline, showing that second order statistics alone are insufficient to produce high performing models. This insufficiency led us to examine the differences between the statistically-matched filters and those from the original supervised model. In this process, we found two additional constraints that, when taken together with second-order filter matching, do produce higher-performing models.

3.3.2 Sparsity in the frequency domain

We found that filters optimized to match only the second order statistics of the fully supervised filters lacked specificity for high-frequencies in the spatial frequency domain, unlike the original target backpropagation filters (Fig. 3-3). We chose to remedy this by optimizing for sparsity in the spatial frequency domain. By making use of the discrete Fourier transform (DFT) matrix [151], optimizing for sparsity in the frequency domain amounts to minimizing the L1 norm of the product of the DFT matrix and each filter. We compute the L1 norm for each filter channel individually and minimize the sum across all channels. Specifically, we sought to minimize:

$$\mathbf{Sparse}(X) = \sum_{f=1}^{nf_i} \sum_{c=1}^{nc_i} |DFT(X_{f,c})|. \tag{3.1}$$

3.3.3 Correlated channels

With backpropagation-trained supervised models, we noticed that each filter generally converges to comparatively similar patterns across its channels. This observation is particularly apparent in layer 1, where filters either are effectively gray (near perfect correlation across all 3 color channel) or are color-opponent (correlation near 1 or -1). However, filters matched to have second order backpropagation statistics and sparsity in the frequency domain do not have such high between-channel correlations (Fig. 3-3). Therefore, we have added a constraint to impose this property. Specifically, for a given layer, we sought to maximize, for each layer i,

$$\mathbf{ChannelCorr}(X) = \sum_{f=1}^{nf_i} \sum_{c=1}^{nc_i} \sum_{d=c+1}^{nc_i} |corr(X_{f,c}, X_{f,d})| \tag{3.2}$$

where X is the filter block tensor (of shape (nf_i, nc_i, fs_i^2)).

3.4 Statistic-Matching Procedure

We wished to evaluate performance of models that had been matched for the three summary statistics described above, comparing them to the performance both of the original supervised back-propagation models as well as random-filter controls.

3.4.1 Sequential Substitution

Our strategy was to start with a fully supervised model, with filter blocks $b_1, b_2, \ldots, b_k$, produce filter blocks $m_1, m_2, \ldots, m_k$ that matched the statistics of the backpropagated filters, substitute these matched filters into the original architecture, and then test the performance of this architecture on the original ImageNet classification task. This substitution was done in a sequential fashion, as opposed to all at once. This was necessary because the summary statistics that we use are agnostic to the ordering of filters in the filter block, and random permutation of the filters would produce the same summary statistics. However, subsequent layers are not invariant to the ordering of the filter block, because it is transposed into the channel dimension at the next layer. Thus, the first matched filter block, m_1, will be out of alignment with the channels of the second back-propagated filter block b_2. Thus, having matched statistics for m_1, we retrained all subsequent layers on m_1, using backpropagation. This procedure produced a new b_2', whose statistics we matched producing a filter block m_2. We then retrained layers 3 and above to produce a target b_3', and so on. This procedure was repeated for all model layers.

3.4.2 Optimizing for summary statistics

To produce the matched filter blocks each layer, we optimized parameters within the filter block so that all three of the summary statistics described above would be matched between m_i and the target supervised blocks. To do so, we linearly combined the three terms (e.g., matching second-order statistics, minimizing DFT sparsity, maximizing channel correlation) into a single optimization criterion:

$$Opt(X) = ||M_i(F) - M(X)|| \; + \lambda_1 \cdot \mathbf{Sparse}(X) - \lambda_2 \cdot \mathbf{ChannelCorr}(X) \quad (3.3)$$

where λ_1 and λ_2 are two hyperparameter weightings. We used the same hyperparameter weightings for all experiments in this work. We found that the performance of the resulting filters is insensitive to the exact hyperparameter choices in a wide range. Because all of our summary statistics are differentiable, we analytically derived gradients and minimized $Opt(X)$ using the L-BFGS-B algorithm [109].

3.5 Are the summary statistics sufficient to produce high-performing models?

To test the extent to which the summary statistics are sufficient to produce high-performing models, we performed the sequential substitution analysis described in the previous section on an architectures whose filers had been trained using error backpropagation [152] on the 2013 ImageNet challenge-set. Unlike the architectures in [6, 127], we simplified our architecture based partly on previous work [127, 153] examining which architectural aspects affected performance. We did not use any normalization layers as we observed, consistent with previous results [153], that normalization layers do not substantially affect performance.

Specifically, we used a 3-layer architecture without intermediate fully connected layers, only including the final fully connected classifier layer. This architecture contained 3 layers consisting of convolutional followed by max-pooling. The filter sizes of the convolutional layers were 7, 5, and 3 square pixels, respectively. The window size of the pooling was 3 square pixels and the stride was 2px. We performed our tests for two versions of this architecture, on in which all layers contained 48 filters, and on in which all layers contained 128 filters. All models were given 128×128-sized images inputs and were trained using the backpropagation procedure described in [6] (ex. batch sizes of 128 images, and a learning rate that was stepped down manually three times when training error decrements plateaued).

For the first two layers of substitution, the performance of the resulting models was indistinguishable from that of the original supervised, backpropagation models, to within error bars. Visually, the filters of the first layer are also very qualitatively similar to the filters of the supervised models (Fig. 3-3). Fig 3-4 and Table 1 present performance results as we sequentially substituted statistically matched filters for layers 1 through 3. For the first two layers, the statistically matched filters are able to maintain the original backpropagation model's performance, suggesting that they are a sufficient summary with respect to producing high-performing models. There is, however, a noticeable loss in performance as the final third layer is substituted into the model, indicating that additional constraints need to be brought to bear to fully summarize the relevant properties of high-performing filters.

3.6 Generalization to other datasets

To quantify the extent that our statistically matched models generalized to other datasets, we measured performance on the Caltech 101 and 256 datasets (Table 2). We found that the substituted filters generalized to the same extent as the original supervised filters. That is, for layers 1 and 2, the statistically matched filters performed as well as the supervised filters. For the third layer, the statistically matched filters had lower performance numbers than the supervised filters, consistent with the lower performance values also observed on the ImageNet challenge-set (Table 1). We measured all performance numbers using a bootstrapped, cross-validated maximum correlation classifier (500 held-out test images) on the outputs of the fully substi-

Model	ILSVRC 2012
3-48-**b**L1-**b**L2-**b**L3	**69.6 ± 4.8%**
3-48-mL1-**b**L2-**b**L3	73.9 ± 4.1%
3-48-mL1-mL2-**b**L3	71.4 ± 4.1%
3-48-mL1-mL2-mL3	79.6 ± 3.5%
3-48-random	93.3 ± 2.2%
3-128-**b**L1-**b**L2-**b**L3	53.1 ± 4.2%
3-128-mL1-**b**L2-**b**L3	**48.4 ± 4.4%**
3-128-mL1-mL2-**b**L3	55.6 ± 5.2%
3-128-mL1-mL2-mL3	68.4 ± 4.3%
3-128-random	89.3 ± 2.7%
3 conv. + no fully conn. layers [127]	71.3%
3 conv + 2 fully conn. layers [127]	**45.4%**

Table 3.1: ILSVRC classification error

Layer names are denoted by ⟨layer number⟩-⟨number of channels⟩-⟨backprop (b) or matched (m) statistics⟩L⟨layer number⟩. For example 3-128-**b**L1-**b**L2-**b**L3 corresponds to a fully supervised model (blue line in Fig. 3-4) and 3-128-mL1-mL2-mL3 corresponds to a model where all filters are derived from summary statistics (light blue line in Fig. 3-4). Also shown are previously reported accuracies from models with similar architectures to ours.

tuted models (the models 48-mL1-mL2-mL3 and 128-mL1-mL2-mL3 in Table 1). We additionally included models of the same architecture with random filters as a control.

3.7 Discussion

These results are lower bound for the extent to which these summary statistics can maintain performance. There are several sources of error which, if fixed, might lead to higher performance with the three summary statistics we use here. First, the backpropagation procedure we use is particularly susceptible to getting stuck in local minima when initial layers of the model are held fixed. For example, as a control, we started with a fully supervised 3-layer model, but reinitialized and re-learned just the third layer filters and the fully-connected classifier layer, while holding the first and second layers fixed. This yielded final models that were significantly less high-performing than the original. This indicates that if we used a backpropagation procedure that was better able to handle just learning top layers, the target statistics

Layer	Caltech 101	Caltech 256
1-48 backprop	47.0 ± 1.8%	19.9 ± 0.66%
1-48-matched-stats	**51.7 ± 0.33%**	**20.6 ± 0.91%**
1-48-random	44.7 ± 1.6%	15.3 ± 2.3%
2-48-backprop	**57.9 ± 2.2%**	**25.6 ± 0.75%**
2-48-matched-stats	53.9 ± 0.75%	22.3 ± 0.33%
2-48-random	50.0 ± 1.5%	18.0 ± 1.3%
3-48-backprop	**65.6 ± 1.3%**	**33.3 ± 0.96%**
3-48-matched-stats	62.4 ± 2.5%	29.5 ± 0.81%
3-48-random	46.2 ± 0.33%	18.3 ± 1.2
1-128 backprop	52.0 ± 0.28%	**21.2 ± 1.2%**
1-128-matched-stats	**52.9 ± 0.74%**	21.0 ± 0.49%
1-128-random	43.1 ± 2.4%	15.2 ± 0.59%
2-128-backprop	**61.1 ± 0.28%**	29.5 ± 1.1%
2-128-matched-stats	59.1 ± 0.38%	**30.1 ± 0.75%**
2-128-random	50.2 ± 1.9%	19.2 ± 1.8%
3-128-backprop	**68.7 ± 1.6%**	**37.4 ± 0.16%**
3-128-matched-stats	58.7 ± 0.84%	27.8 ± 1.4%
3-128-random	50.0 ± 2.1%	19.7 ± 1.6
4 conv. layers (unsupervised) [15]	**71.0 ± 1.0 %**	33.9 ± 1.1%

Table 3.2: Caltech classification accuracy

Layer names are denoted by ⟨layer number⟩-⟨number of channels⟩-⟨optimization type⟩. The "matched-stats" entries represent outputs from fully substituted models (all filters are determined from matching the backpropagation summary statistics).

learned for the higher layers in the sequential procedure described in section 3.4.1 might be of better quality. Secondly, for the larger 128-filter model, the backpropagation procedure takes many epochs to converge, substantially increasing the amount of time necessary to re-train the model multiple times. We might have had modestly better performance if we ran the backpropagation procedure for the 128 channel mL1-mL2-bL3 model (brown line of Fig. 3-4) until performance plateaued. However, even with these factors, we still suspect that additional statistical constraints will be necessary to fully capture the performance of higher-level layers.

It has been observed that in supervised models trained with backpropagation, especially with weight-decay, that a significant fraction of filters are "dead", meaning that across the spatial and channel dimensions, the filter has very low variance. Unlike the backpropagation models, the statistically matched filters did not have dead filters — a seemingly desirable property as minimal, if any, useful information passes through dead filters. Another observable difference between the statistically matched and backpropagation filters is that the latter do not show the rough circular aperture seen in most of the backpropagation filters (Fig. 3-3). The significance of the lack of an aperture is unclear beyond layer 2 where the filter size is small, but for the first two layers our substitution analyses indicate that this difference does not appear to affect performance.

3.8 Conclusions and future directions

We have found that matching three simple statistical summaries of filters learned by backpropagation algorithms can be useful in producing performance gains over random filter controls. The performance gap between the statistically-matched and original fully-supervised filters is negligible for the several network layers, but increases thereafter. Each of the three statistical summaries on their own are unlikely to be successful at producing above-random performance, but taken together they constrain filter values in a meaningful way.

This work will have value for producing better unsupervised training algorithms if several key obstacles can be overcome. One such obstacle is that the three statistics that we match here are, obviously, insufficient at higher network layers. Following our approach, remedying this problem will involve identifying additional statistical summaries that can be matched. Going forward, we will compare statistically-matched filters we have built so far with those from fully supervised networks, at higher layers, to determine statistical divergences between them.

A second obstacle is that the statistics we use here are determined from fully-supervised filters. Producing a fully unsupervised procedure would involve showing how to learn these statistical summaries from the original input data. However, as we saw in Fig. 3-2, it appears that there is a strong relationship between the natural image statistics and filter statistics that we compute. We suspect that this relationship can be stated in the form a mathematical transform that produces filter correlation statistics as a function of the input correlation statistics. In this work we have worked under an ideal case where we have such a transform (by using the "transform" empirically found in high-performing models). This assumption allowed us to test the sufficiency of statistical constraints and would have otherwise been confounded if we tested both the sufficiency of a transform and our statistical constraints simultaneously. One potential route to understanding this transform is via recent analytical work in deep linear networks [154], where it is shown for linear networks that the weight update equation is explicitly determined by a combination of second order image statistics and image to label statistics. The conclusions of that work do not directly apply here because the networks we use include a nonlinear max-pooling operation. However, given a fixed set of max-value switches (as discussed in [15]) for a set of training images, the networks we work with here are linear. In future work, we will explore this idea more fully.

A final obstacle arises from the fact that we have used comparatively simplified architectures here to test our ideas, e.g smaller numbers of layers, no point-wise ReLu nonlinearities, and no normalization operations. Because state-of-the-art architectures do use these more complex operations and are deeper, we will have to show that the ideas here transfer to those larger networks before a state-of-the-art fully unsupervised procedure can be attained.

Even with procedures differing entirely in approach (different formulation of objective functions), the second order statistics described here may provide utility as a quick check-sum for estimating a model layer's performance—layer 1 second order statistics of high-performing models often converge to similar patterns on the same dataset even when the model architecture differs. Therefore, if one has a high-performing model on a dataset (ex. trained via backpropagation) and wishes to quickly test the performance of a model generated by an unsupervised learning procedure, the correlation of the model in question's second order statistics with the high-performing model's second order statistics may give a quick sense of whether the model in question will perform well on that dataset.

More broadly, we suggest a general approach for further refining and understanding the statistical properties sufficient to produce high-performing models: (1) given a set of relevant summary statistical constraints, like those presented here, optimize to find filters matching the summary statistics of a high-performing (supervised) model; (2) then compare the optimized filters with the original supervised filters and characterize any systematic differences between the two; (3) once characterized, formalize this difference into an additional statistical constraint and repeat. With this loop, model performance is checked to prevent from including constraints that are orthogonal to model performance and arise from potential artifacts of the supervised learning procedure.

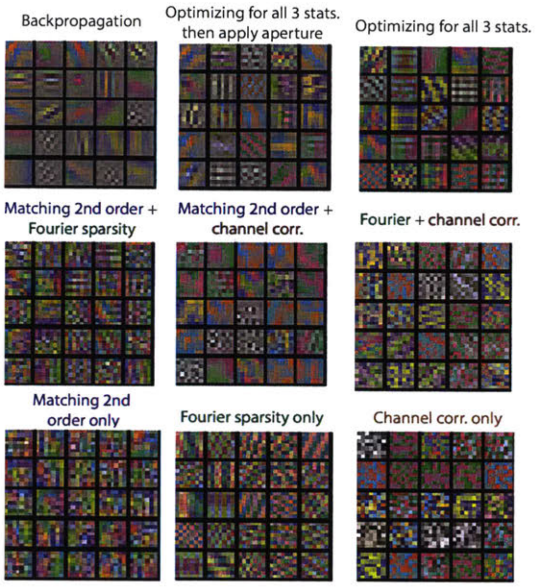
Backpropagation
Optimizing for all 3 stats. then apply aperture
Optimizing for all 3 stats.
Matching 2nd order + Fourier sparsity
Matching 2nd order + channel corr.
Fourier + channel corr.
Matching 2nd order only
Fourier sparsity only
Channel corr. only

Figure 3-3: Filter visualizations
(Row 1, Col. 1) Filters from a 128-channel backpropagation model. Shown are the 25 filters with the highest variance. A relatively large proportion of the filters in this model and others trained via backpropagation are "dead" (have low variance—appear uniformly gray). **(Row 1, Col. 2)** Filters found by optimizing for the aforementioned summary statistics (second order statistics, sparsity in the frequency domain, correlated channels) of the 128-channel model shown above. There are two main differences between these filters and the original backpropagation filters. (1) There are essentially no dead filters and (2) the absolute magnitude of filter values (summed across all filters and channels) is not localized towards the spatial center of the filters (i.e., the oriented gratings of the backpropagation filters are almost always centered spatially within the filter and surrounded by a rough circular aperture of gray). At least for layers 1 and 2 of the convnets tested, this aperture is not critical for performance. For visualization, we apply the same aperture to our statistically matched filters. **(Row 1, Col. 3)** Filters from **(Row 1, Col. 2)** but without the aperture. The filters without the aperture were used in our substitution experiments. **(Rows 2 and 3)** Filters optimized from choosing two out of three of our aforementioned summary statistics—in all cases, a key aspect seen in the backpropagation filters is missing in these filters (ex. frequency selectivity **(Row 3, Cols. 1 and 3; Row 2, Col. 2)**, consistency across channels **(Row 3, Col. 2; Row 2, Col. 1)**, low frequency filters **(Row 2, Col. 3)**).

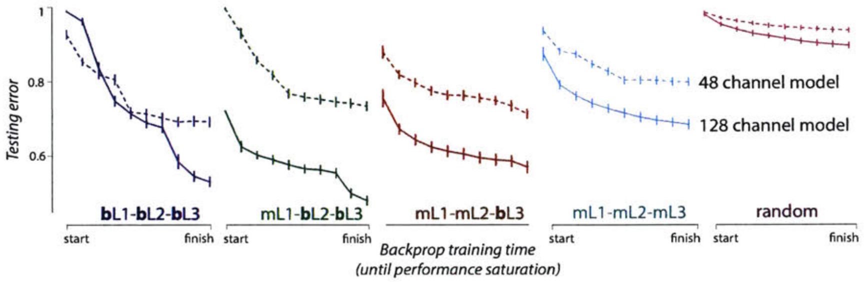

Figure 3-4: Substituting in filters with matching summary statistics
Dashed lines represent substitutions into a 3-layer 48-channel model and solid lines represent substitutions into a 3-layer 128-channel model. The blue lines represent the test curves on ImageNet as each model is trained via backpropagation. At the end of the backpropagation training of the blue lines, layer 1 filters optimized to have matching summary statistics of the backpropagation-trained models are substituted in. While the new, substituted layer 1 filters are held fixed, layers 2, 3 and the fully connected classifier layer are re-trained via backpropagation (green lines). The same procedure is repeated again: layer 2 is now replaced with filters matching the backpropagation-trained summary statistics and then held fixed as layer 3 and the fully connected layer are re-trained (brown lines). After, the same procedure is performed again with layer 3 substituted, resulting in a model containing only filters derived from summary statistics (light blue line). As a control, we also show the test curve of a model containing only random filters while training the final fully connected layer on top (pink lines). In each plot we show the 48-channel and 128-channel models on the same x-axis for ease of visual comparison. However, the 128-channel model has a slower backprop convergence in terms of the number of epochs. We binned each original time-course into 10 equal blocks and plot the average performance within each block (bars indicate SEM within the block).

Chapter 4

The visual object representation of juvenile primates

Visual object recognition is a computationally challenging task that is critical to everyday functioning. In human and non-human primates, object recognition relies on information processing along the ventral visual stream and the resulting neural population representation at the top of that stream — the inferior temporal cortex (IT). Current neurally-mechanistic computational models of this system (in particular, artificial neural networks, ANNs) can explain and predict much of the image-evoked activity of the neurons along the ventral stream up to and including IT and the resulting behavior of the subject. These ANN models successfully approximate the rapid (<200 msec) flow of spike rate information along the adult ventral stream, but they tell us nothing about how that successful processing stream is set up. Put another way, what are the developmental and post-natal learning mechanisms that are critical to the performance of the ventral visual stream? Because the aforementioned models start from an initial macro- and meso-architecture but their high correspondence to the adult ventral stream gradually improves with supervised "training" on object categorization tasks, one hypothesis is that the initial ANN architecture corresponds to the work of genes and developmental mechanisms, while some biological plausible version of model training corresponds to primate post-natal visual learning. However, this hypothesis and many others cannot yet be engaged because we know next to nothing about the neurophysiological status and capabilities of higher levels of the ventral stream in either newborn or juveniles non-human primates. Here we[1] performed the first large-scale multi-electrode recordings (~200 total sites) of spiking neurons in IT of three juvenile (~25 weeks post-natal) awake behaving macaques, and we compare and contrast these with >900 recording sites obtained using identical methods in adult IT and V4. We used a broad range of comparison metrics, including — most critically — population-level metrics that assessed the ability of the IT population to produce high performing object categorization behavior and explain primate behavioral patterns over different object tasks. Our most striking finding was that, by all such population measures, juvenile IT is statistically indistinguishable from adult IT, even though our methods have the power to distinguish the IT representation from its dominant input representation (V4). Interestingly, we also found that the visual response latency is approximately 50 ms slower in juvenile IT. In sum, our results show that the ventral visual stream already presents a highly performant, adult-like object representation as early as 25 weeks of age, and this argues that either pre-natal development plays a powerful role in setting up this system and/or that the post-natal visual learning is rapid.

4.1 Significance

Little is known about how about the postnatal development of inferior temporal cortex, thought to subserve visual object recognition. While prior fMRI studies have provided a spatially coarse time-line of activation and associated developmental properties, electrophysiological studies have been less common in higher-visual cortex. Existent electrophysiological studies have used stimulus sets which have not fully engaged core object recognition abilities, precluding the use of population-level readouts. Here, we utilized array-based physiology in macaque inferior temporal (IT) cortex to collect large-scale datasets in animals aged 19-32 weeks, and have bench-

[1] Work done in collaboration with Najib Majaj, Kohitij Kar, Lynne Kiorpes, J. Anthony Movshon, and James DiCarlo

marked this data against >900 multi-unit sites in adult IT and V4. Consistent with prior electrophysiological studies, we observe, at 19-32 weeks, visual response latencies approximately 50 ms slower than any of our adult recordings. However, we do not observe any deficits in our measures of neural categorization performance, or behavioral consistency measures, suggesting that, even at half a year of age, IT already supports a robust representation capable of supporting challenging recognition tasks.

4.2 Introduction

View-invariant object recognition is a computationally difficult but highly relevant behaviorally cognitive task. Retinal images of real-world objects vary drastically due to changes in object pose, size, position, lighting, non-rigid deformation, occlusion, and many other sources of noise and variation. Humans effortlessly recognize objects rapidly and accurately in spite of this enormous variation, an impressive computational feat [155]. This ability is supported by a set of interconnected brain areas collectively called the ventral visual stream [156, 157], with homologous areas in non-human primates [158, 159]. The ventral stream is thought to function as a series of hierarchical processing stages [160, 161, 162], that encode image content (e.g. object identity and category) increasingly explicitly in successive cortical areas [163, 155, 2]. For example, neurons in the lowest area, V1, are well-described by Gabor-like edge detectors that extract rough object outlines [1], though the V1 population does not show robust tolerance to complex image transformations [2, 8]. Conversely, rapidly-evoked population activity in top-level inferior temporal (IT) cortex can directly support real-time, invariant object categorization over a wide range of tasks [4, 5, 8]. Mid-level ventral areas — such as V4, the dominant cortical input to IT — exhibit intermediate levels of object selectivity and variation tolerance [5, 81, 164, 8].

From past developmental studies, three main hypotheses can be formed regarding the development of postnatal visual object recognition, as described below. Because no known prior study has explicitly tested neural recognition performance at an early age, support for each of these hypotheses is indirect and at most suggestive.

Hypothesis 1: neural responses in high-level visual areas are adult-like at birth or within several weeks after it. Anatomical and eletrophysiological studies in early visual areas often find that infants reach adult-like states very early on in development. Connections from the Lateral Geniculate Nucleus (LGN) to V1, for instance, are adult-like by 8 weeks [165]. Retinotopic maps are generally considered the first stage of development in V1 and develop without retinal inputs [166], and orientation selectivity develops following retinotopic organization [167]. Cortical projections to IT appear adult-like by 7 to 18 weeks [29] and projections from IT to parahippocampus and perirhinal cortex appear adult-like at one week. A caveat is that many anatomical studies use qualitative approaches to determine if projections or areas are adult-like, and may not be sensitive to more subtle developmental differences.

Hypothesis 2: neural object recognition performance is at or quickly within adult-levels several weeks to months after birth, despite neural responses being immature with other measures (such as visual response latency). Prior work has demonstrated longer visual response latencies in animals ranging from 4 to 28 weeks of age with adult-like single-site selectivities [30, 168]. However, these studies showed a limited number of cartoon-like stimuli which are insufficient to test the full capacity of invariant object recognition supported by primate vision. Aside from these results, there are no other known experiments recording IT responses in awake infant or juvenile macaques. While not directly suggestive of invariant object recognition, human studies have provided evidence of object permanence existing in 20 week old infants (approximately 5 week old macaques) [169], contrary to earlier work suggesting later development [170].

Hypothesis 3: higher visual cortex gradually becomes refined and more optimal at recognizing objects months to years after birth. Recent work has shown face selective patches, as measured by fMRI, do not appear to develop in face deprived macaques [39], perhaps not unlike the abnormal developmental trajectories observed in early visual areas when given altered visual experiences [171, 172]. The face patch findings suggest that some experience-dependent organization occurs during development of higher visual areas [39], although the exact nature of this organization and its relationship to both neural selectivity and population-level decoding remains unclear. Furthermore, normal visual experience may be generally necessary to maintain organization rather than to create it [28]. Human fMRI studies have found face patches in the earliest tested ages (12-32 weeks, approximately 3-8 weeks in macaques), but found differences from adults using representational similarity measures [37]. Learning experiments in adult macaques show significant neural plasticity in response to visual experience [21, 20], however the relationship between adult plasticity and development are unknown. Additional lines of work from psychophysical studies often suggest a gradual emergence [40, 28] of visually-guided behavior on the scale of approximately 50 to 300 weeks depending on the task [28]. These tasks have assessed sensitivity to spatial resolution, visual acuity, global form, pattern motion, contour integration, among other properties [28, 173, 174, 175, 176, 177]. Behavioral studies inherently do not speak to where development is occurring, leaving open the possibility that decision-making or motor areas may be bottlenecks to performance rather than the underlying visual representation itself [28].

To investigate juvenile IT and assess how, and if, it differs from the adult state, we sought to collect data-sets to allow us to compare developmental responses to adult responses using a battery of tasks and metrics used previously to establish links between neural responses and behavior [8, 10]. More specifically, we sought to benchmark developmental IT using population-level metrics assessing neural categorization performance, behavioral consistency, and representational similarity, along with single-site metrics (such as latency, and d-prime measures). To support this diverse array of testing, we used a challenging image-set (Figure 4-1) consisting of thousands of images previously used to expose key computational difficulties in visual object recognition [8, 10]. We showed these images using a rapid-serial-visual presentation paradigm to adult and juvenile macaques (aged 19-32 weeks). While animals viewed the images, we recorded electrophysiological responses to 96-channel Utah arrays implanted in pIT and V2. Fairly performing cross-animal analyses required controlling many experimental factors. We, therefore, developed a bootstrap procedure to carefully match trial, image, and neural sites across animals to remove these nuisance factors.

While we observe our juvenile pIT samples exhibit visual response latencies 50 to up to 100 ms slower than our adult samples, we do not observe such large deviations in other metrics. Juvenile categorization performance, for instance, is within the range of our adult IT samples. Although our juvenile data exhibit more latent responses than adults, the neural responses between 100-150 ms still represent sufficient information to accurately classify categories at the level of our adult samples. Our results provide evidence for IT supporting high-levels of categorization performance, in a manner representationally indistinguishable from adults, on challenging object recognition tasks as early as 25 weeks.

4.3 Materials and methods

4.3.1 Image-set generation

We used a subset of the image stimuli previously reported [8]. These images were generated from ray-tracing software (http://www.povray.org). The stimuli were designed to expose key aspects which make object recognition computationally difficulty. For this reason, each image was generated to depict one of 64 objects at a random size, location, and orientation in the image. To de-correlate the background from the object, random backgrounds depicting natural scenes were shown in each image. Each object belonged to one of 8 categories and each category contained 8 objects (animals, boats, cars, chairs, faces, fruits, planes, tables). For example, the cars category contains 8 separate car models as objects.

The stimulus set was designed to provide a range of difficulties with respect to object recognition tasks. Images depicting objects in a canonical view in the center of the screen were psychophysically easier to recognize than objects in highly eccentric angles near the edges of the image. Whereas, objects near the edge of the screen in highly eccentric positions were much harder to recognize. Essentially, the level of "variation" of object size, position, and location controls the difficulty of recognizing

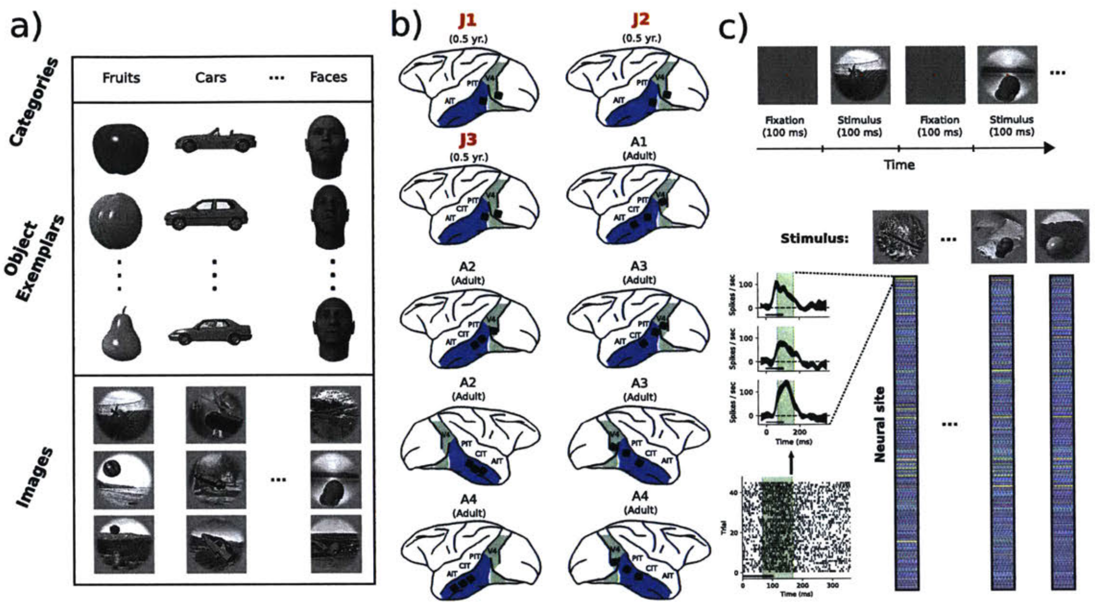

Figure 4-1: Experimental design

a. Our stimulus set contains complex three-dimensional objects at high levels of position, pose, and size variation, and placed on natural backgrounds. Recordings for the images have previously been made in adult macaques and shown to expose key aspects of invariant object recognition that differentially engage inferior temporal (IT) cortex [8, 10]. **b.** We have placed two 96 channel arrays: one in V2 and one in posterior IT (pIT) in three animals (at 19, 25, and 26 weeks of age). We additionally have utilized adult IT and V4 data collected from prior published and unpublished studies [8]. **c.** We showed stimuli using a rapid-serial visual presentation (RSVP) paradigm. Animals fixated at a small (0.25°) red fixation point at the center of the screen and images were shown in pseudo-random order for 100 ms followed by a blank screen (still containing the fixation point) for 100 ms. This sequence was repeated either until 8 images were shown or fixation was broken. We showed each image up to 45 times. For many of our analyses, we average the responses of each neural site (across trials) to each image to create a stimulus response vector. Stimulus response vectors contain responses of all neural sites (in a given brain or age-defined area) to each stimulus image in the interval of 70-170 ms. In some analyses, we use alternative time windows for construction of these vectors.

the object. The stimulus set was divided initially into three subsets (variations 0, 3, 6). At variation 0, the objects are all shown in a canonical view centered in the image. At variation 3, the objects are rendered in positions, sizes, and orientations that fell within the ranges: *x*: [$-1.2°$, 1.2°], *y*: [$-2.4°$, 2.4°], *pose*: [$-45°$, 45°], and *size*: [$\times\frac{1}{1.3}$, $\times 1.3$]. Variation 6 used an increased range of positions, sizes, and orientations: [$-2.4°$, 2.4°], [$-4.8°$, 4.8°], [$-90°$, 90°], and [$\times\frac{1}{1.6}$, $\times 1.6$], respectively. Images were rendered at 256 × 256 px in gray-scale.

We chose to focus our data collection efforts on the variation 3 and 6 datasets because we hypothesized that it would more readily allow us to separate between representations of differing underlying recognition abilities. Variation 3 and 6 both contain 2560 images (64 objects shown 40 times with randomly assigned latent parameters from the ranges described above). For some of the adult arrays we further focused collection efforts around a combined subset of variation 3 and 6 (containing 320 images from each, and 8 distinct objects) — we refer to this reduced set as variation 640.

4.3.2 Neural data collection

We used three male Macaca nemestrinas (all between 19-32 weeks old at the time of recording) for our developmental studies. We trained the animals for several weeks using juice rewards and operant conditioning, acclimating them to our rapid-serial-visual presentation paradigm. Unlike prior studies in adults which use surgically implanted head posts, we chose to use custom head masks to stabilize their head positions similar to [178]. These masks were used to reduce the invasiveness of our experiment and because they have had a long history of use in younger animals.

We monitored eye position using the SR Research Eyelink II video tracking system. Animals calibrated by fixating on a red fixation point (0.25°) shown at positions on a computer monitor. During the main experiment, animals fixated at the red square while images (typically 8) were shown in rapid succession (100 ms on, 100 ms off, subtending 8° with a resolution of 32 pixels/°) overlaid on a gray background. Animals were rewarded with juice and a tone at the end of each trial if fixation was maintained within ±2.5° of the red square for the entire sequence of 8 images. If fixation deviated from this window during the trial, the trial was aborted and the neural responses during that interval discarded. We showed stimuli in a random order across multiple days. Each image stimulus used in our analyses was shown a minimum of 23 times (typically ~45).

Following prior studies, we placed arrays (Figure 4-1) in posterior IT and V2 as guided by sulcus patterns, visible during surgery and in our pre-surgery anatomical MRI scans. All surgical procedures and recordings for our developmental data were performed at New York University and were approved by the University Animal Welfare Committee. Because of technical hardware limitations, we were not able to record from both the IT and V2 arrays simultaneously for all of our experiments. We used most of our recording time for the IT arrays collecting the image-sets descried above. For the V2 array we used most of the recording time for other non-overlapping image-sets used in unrelated experiments; consequently, we focus entirely on our IT recordings here.

In our analyses, we included data collected in four adult male animals. All animals were trained and recorded from using the same procedures above, with the exception of using surgically implanted head posts for head stabilization instead of the custom masks we used for the juveniles. All surgical procedures and recordings for our adult data were performed at MIT and were approved by the Committee for Animal Care. Data from two of the adult animals have been previously reported in [8] and contained 2 arrays in IT and 1 array in V4 per hemisphere (for one monkey in this set, we have recorded separately from both hemispheres using the same array placements). Data from the remaining two adult animals were obtained as part of an unrelated study and have not currently been reported elsewhere. These two monkeys were implanted each with three arrays in IT. In a separate surgery, one of the monkeys was further implanted with a V4 array in the other hemisphere. We have also collected behavioral responses of these later two monkeys performing a recognition task which we use for our behavioral similarity measure also reported elsewhere [179]. All array data were collected over the course of several months (typically 2 months or more) for each animal.

4.3.3 Thresholded spike events

All analyses are performed on binned threshold multi-unit events. All data were first band-pass filtered (250 Hz to 7.5 kHz) and sampled at 30 kHz using Blackrock Neural Signal Processors. Thresholds were set for each channel on each day (and in some cases more than once per day) as -3× the RMS of each channel's background activity while the animal was staring at a uniformly gray background screen. A threshold event was defined as a channel's voltage (falling edge) crossing -3× the RMS of the baseline voltage.

4.3.4 Unit selection

Units were selected based on visual drive estimated on an independent stimulus set [8]. Visual drive was defined as the d-prime between a unit's cross-validated (across trials) highest evoked responses (top 10% of image stimuli) vs. blank screen presentations. Image bootstrapping (while scrambling the "image" and "blank" stimulus labels) was used to bootstrap a null distribution of visual d-primes and set a threshold of inclusion for each unit (>80% of the image scrambled bootstrap visual d-primes). We further screened neural sites based on their trial reliability. We defined the trial reliability for each neural site as the mean (bootstrapped) Pearson correlation between two image response vectors estimated from random, non-overlapping trial subsets. Each element in an image response vector represents a neural site's trial-averaged spike count from 70-170 ms post-stimulus onset in response to a particular image, with other elements representing other image responses. Similar to the visual d-prime measure, we also estimated a null distribution for each neural site by scrambling image labels (for each neural site, one of the two image response vector's elements were scrambled) and used this distribution to set a threshold for each unit's inclusion (>99.5% of the image scrambled reliability).

For the variation 0 and 3 datasets, our unit selection procedure (which was performed once on a single independent image set) yielded 196 units for our juvenile pIT pool, 171 units for our adult V4 pool, 141 units for our adult pIT pool, and 143 units for our combined cIT and aIT pool. For the variation 640 dataset we had 196 units for our juvenile pIT pool, 308 units for our adult V4 pool, 189 units for our adult pIT pool, and 449 units for our combined cIT and aIT pool. Some adult animals were not shown the full variation 0 or 3 datasets, resulting in us having more neural sites for the variation 640 dataset.

4.3.5 V1-like model

We chose to include a control model representation to provide context and help provide grounding for our neural data and metrics. A V1-like model, we hypothesized, should also provide something of a lower bound for performance-based metrics which we would expect all of our V4 and IT samples should out-perform. We used the V1-like model introduced in [36], and previously used as a benchmark in similar contexts [8, 10]. Briefly, the model takes as input a set of pixel inputs for each image, and outputs a feature vector (˜76 thousand visually-drivable features which we restricted our analyses to) for each input image. Each feature dimension was processed in subsequent metric evaluation analyses as if it were a "neural site" which provided us with responses to each stimulus image. Processing the data in this way allowed us to compare the V1-like model to the neural data using the same metrics and procedures.

4.3.6 Metric evaluation methods

The general overview of procedures is: (1) pre-processing and unit selection, (2) bootstrap sampling, (3) metric evaluation with the sampled data.

4.3.7 Pre-processing

For all population-level metrics (i.e., not the single-site metrics in Figure 4-2), the following paragraph applies. Aside from this paragraph, the remainder of steps in this section apply to both single-site d-prime and population-level metrics. We first subtract each neural site's response to blank images from the stimulus image trials. We subsequently normalize each neural site by dividing each site's image responses by the standard deviation across images.

For each trial and neural site, we then compute the average binned response over the 70-170 ms time interval (unless stated otherwise in our analyses specifically relating to time). For each stimulus block, we then subtracted from each neural site's responses, the site's mean image-evoked activity across all stimuli within the stimulus block. A stimulus block is defined as the set of images shown within a single experimental run in pseudo-random order.

Finally, each array is independently common mode (average response to each image across all units) subtracted. In analyses in which the metric is corrected based on trial-trial reliability (the behavioral consistency and representational dissimilarity metrics), the common mode subtraction step takes place on each trial splitting independently, then the trial-trial reliability is estimated. The final step of pre-processing consists of trial averaging each neural site's responses across image stimuli — for each unit, a 2D matrix is emitted of image responses by neural sites (see Figure 4-1).

Except for RDM analyses, all data are trial matched and additionally for performance metrics, bootstrapped over train/test splits. RDM analyses are not trial matched but are noise-corrected based on split-half trial-reliability. When evaluating a pool (ex. the pool of adult pIT data) with more units than the analysis calls for, unit selection is also bootstrapped. When evaluating a pool with more trials than called for, trials are also bootstrapped.

4.3.8 Bootstrap sampling for population-level metrics

We start with the total pool of all arrays we have collected for a given brain area and age range (ex. all adult pIT arrays). The next step is to select a sub-pool of arrays containing a sufficient number of trials of the stimulus set to be analyzed (for some arrays, we only have data on the variation 640 images, so these arrays would be excluded here if we were analyzing responses to variation 3 images).

The next step is the random selection of 2 arrays from the pool of valid arrays.

This step allows us to avoid relying on the assumption that each neural site in an array represents a completely random sample of, for example pIT. Two such factors which may make this assumption invalid include: spatial sampling (pairs of neighboring electrodes on an array are approximately 0.5 mm apart and may exhibit similar visual responses due to spatial proximity), and cross-hemisphere variance where neural responses within a hemisphere may exhibit higher similarity than when compared with another hemisphere. We chose to sample 2 arrays in any given bootstrap to simultaneously maximize the amount of data used during each bootstrap, while still leaving out at least 1 array from each pool for the purpose of having the bootstrap procedure capture across-hemisphere variance.

Finally, from the pool of 2 arrays, we randomly select n_{units} and n_{trials} to create a single feature matrix of $[n_{images} \times n_{trials} \times n_{units}]$. This concludes a single bootstrap sample of our data for a given brain or age-defined area. For each of our two main stimulus sets (variations 3 and 640) we chose n_{trials} for the purpose of both maximizing the number of arrays and trials used. For variation 0, 3, and 640, we set n_{trials} to 23, 37, and 23, respectively. Similarly, we chose n_{units} to maximize both the number of arrays and total data used. For variation 0, 3, and 640, we set n_{units} to 58, 58, and 49, respectively.

For RDM analyses, we ensured that when computing the similarity of a brain area to itself (ex. "adult cIT & aIT" RDM correlations to "adult cIT & aIT") data from the same array never resided simultaneously in both the test and target RDMs. This requirement, as mentioned above, mitigates the need for assuming the individual neural sites within a single array represent independent, random neural samples.

4.3.9 Metric-specific methods

Response onset latency and latency adjustment

We defined response latency using a normalized threshold crossing measure. We used the following pre-processing for this analysis. Spike counts were first binned into 10 ms bins for each trial (from 0 ms to 250 ms post-stimulus onset). For each neural site, we then computed two average time series (from 0 ms to 250 ms post-stimulus onset): (1) the mean response to all image trials, (2) the mean response to all blank trials. We then subtracted the time series of blank trials from the image trials. Next, we normalized (zero mean, unit variance) each neural site's time series and then smoothed across time using a Gaussian kernel with a standard deviation of 30 ms. We performed an additional neural site rejection step with two criteria: (1) the maximum of the filtered time series must rise above 1.1 standard deviations at least once in the time series, (2) the first time bin (0-10 ms) must be less than 0.5 standard deviations. We defined the latency of each neural site as the first time bin which rises above 0.5 standard deviations.

For some analyses, we use each neural site's latency to adjust the bins used as input into other metrics (namely, single-site d-prime measures—we do not show latency-adjusted population-level metrics here). The latency of each neural site, as estimated above, is used to define the start of a 100 ms bin, which we then process identically to other neural data estimated from the standard 70-170 ms read-out window. Note that because we reject some additional units when estimating latency, analyses which are latency corrected may have fewer units than analyses using the standard 70-170 ms read-out.

Sparsity

We used a measure of sparseness as defined in [180]. The metric quantifies the extent to which individual neural sites exhibit "sharper" tuning: high responses to few images, as opposed to responding to all stimuli with the same magnitude. Sparseness was defined for each neural site as

$$s = \frac{(\sum_{i=1}^{N} \frac{r_i}{N})^2}{\sum_{i=1}^{N} \frac{r_i^2}{N}}, \tag{4.1}$$

where r_i represents the neural site's response to the i^{th} image and N represents the number of images in the stimulus set.

Single-site d-primes (category, object, and face identification)

These measures provide estimates of each neural site's selectivity. The category d-prime was defined as the d-prime between each neural site's highest responding category vs. its lowest responding category. The highest and lowest responding category were computed on a separate half of trials than those used for the computation of the d-prime.

The object and face single-site d-primes were similarly defined to the category d-prime. Instead of computing the d-prime for the highest and lowest responding category, we compute the d-prime between the highest and lowest responding object (or face).

The variation 3 dataset, for which we compute these metrics on, contains 8 categories with 8 objects per category (each object is depicted in 40 images in random positions and orientations on a random background). Each bootstrap evaluation consists of trial-matching the data (using a random selection of 37 trials).

Face detection d-primes

This measure was defined as the d-prime between all face images and all non-face images. Unlike the face identification d-prime, which measures the ability of a neural site to discriminate between its highest and lowest responding face, this measure quantifies the extent that a neural site detects faces from non-faces.

We used the variation 0 dataset to estimate d-primes. Bootstraps are across a random selection of 23 trials for all data.

Category and object population performance

Similar to the category and object d-primes, the performance metrics provide estimates of neural performance on one vs. all category recognition tasks. Unlike the category and object single-site d-primes, these measures provide estimates of a population's ability to separate categories.

Cross-validated (90% training images, 10% testing) linear support vector machines

($C = 1$ for the variation 3 and 640 datasets) were trained on neural samples with the category identity depicted in each image as the label. Percent correct testing-set classifications were then averaged across categories to provide a population-level estimate of performance.

We present classification performance relative to the performance of humans completing an 8-way categorization task using the same stimuli for which we collected electrophysiological data (we divided our performance estimates by the mean human performance across participants and categories). The human data was previously used to demonstrate the consistency of simple, weighted, linear IT read-outs and human behavior [8]. Human data were collected using procedures approved by the Massachusetts Institute of Technology Committee on the Use of Humans as Experimental Subjects. In total, 29 humans participated in the subset of data we used from [8]. Participants completed 30-45 minute sessions via Amazon Mechanical Turk for a small payment. Trials began with a small fixation point shown for 500 ms, followed by an image appearing at the center of the screen for 100 ms. Following a delay of 300 ms, the participant was given eight images and was tasked with choosing the one which matched the category of the test image. After selecting a response, a new trial began. Participants judged a random subset of 400 images. No feedback was given to the participant regarding the correctness of their responses.

Face identification performance

The methods for face identification were the same as those used for categorization performance with the following exceptions. Cross-validation used 80% training images, and 20% testing images as a consequence of the smaller number of images in the variation 0 dataset used to estimate identification performance (80 images of the 640 total images in variation 0 depict faces, with 10 images per each face identity). We additionally used a higher regularization value to prevent classifier overfitting on the reduced number of training images ($C = 1000$), however, we found performance relatively insensitive to this parameter.

Face probability as a function of distance

To visualize the extent to which face sites were spatially clustered in adult and juvenile pIT, we plotted the probability of finding two face sites as a function of spatial distance. Representations which are more "patch-like" would have higher probabilities of finding pairs of face sites at smaller distances, and this probability would decrease at farther distances (a negative slope). Conversely, representations which are randomly distributed across space with face sites, would show equal probability of finding a pair of face sites regardless of distance (a line with zero slope).

Because aligning arrays across animals and hemispheres presents unique problems, not only due to uncertainties in exact (sub-millimeter) positioning and orientation of array implants but also due to uncertainties regarding animal-animal anatomical variability, we did not attempt to compare across arrays. We performed the analysis only within sites from the same array, and combined the counts at each spatial distance at the last step of analysis.

The analysis itself consisted of simply counting (from each face selective site) the number of face sites found at each pair of spatial distances and normalizing (dividing) this by the total number of visually responding sites at each of the distances from the face selective site. Because our electrodes are arranged in a regularly spaced grid, there were only a small faction of pair-wise spatial distances that existed allowing us to more easily pool over positions (unlike if we had done single-electrode recordings where each penetration would be a unique distance from all others). We defined a face selective site as those which had face vs. non-face d-primes (on our variation 0 dataset) above a threshold defined for each visual or age-defined area. This threshold was determined by randomly shuffling the image labels of the data and computing face vs. non-face d-primes. We chose the 99th percentile of the scrambled d-primes for our threshold—any non-scrambled unit with a d-prime above this would be considered face-selective.

Behavioral similarity

We used an image-level one-versus-all performance metric. This metric, called B.I1n, measures the performance of recognizing each object aggregated across all distractor objects in the stimulus set [179]. The B.I1n vector represents image-level performances on multiple object classification tasks, with each element representing the performance of recognizing a particular stimulus image (see Figure 4-4). B.I1n vectors can be estimated from neural responses, models, and, importantly, behavioral responses.

We estimated B.I1n vectors on the variation 640 dataset which contains 8 objects each depicted in 80 images. Two vectors are used to compute the B.I1 vector (which is then normalized to produce the B.I1n vector): a hit rate for each image (fraction of trials the image was correctly classified) and false alarm rates for each object (across all images). The false alarm rate represents the rate at which an object is reported as being in the test image when it was not actually depicted. With the hit and false alarm rates, each element of the B.I1 vector is defined as the normalized (via the inverse CDF) difference between the hit rate for image i and the false alarm rate for the object depicted in image i (Figure 4-4), resulting in a vector with an element for each image in the stimulus set. Finally, the B.I1n vector is computed by subtracting from each element, the mean d-prime of each object (see Figure 4-4). Normalization helps focus the metric on image-level variance, instead of mean differences between object performances which can dominate the variance of the B.I1 vector.

Estimating B.I1n vectors from monkey behavior The behavioral paradigm we used for collecting monkey behavior utilized eye position for behavioral reports from the animal. Trials were initiated by fixating at a white square dot (0.2°) for 300 ms. A test image (depicted on the far left column of Figure 4-4) was shown for 100 ms and then, after a delay of 100 ms where a blank screen was shown, a canonical view of the object (Figure 4-4) in addition to a canonical view of a random distractor object was shown for up to 1500 ms [181]. The task then was to choose the canonical view that depicts the same object as the test image by holding fixation for 400 ms on the correct object. Before the selection phase, trials were aborted if gaze was not within ±2° of the fixation dot. We collected behavioral responses from two adult monkeys using procedures approved by the MIT Committee for Animal Care [181, 179].

Estimating B.I1 vectors from neural responses and models Each feature sample, either from models or from neural sites is processed identically with respect to the B.I1n and derived metrics. First a linear SVM classifier ($C = 10$) is trained for each task (object i vs. all other objects). Cross-validation is used to extract classifier decision boundary estimates for each held-out test image (5 fold classification with 80% testing, 20% testing images) and classifier. This procedure results in a vector of 8 classifier decision outputs for each image. The softmax function is then used to normalize the 8 values for each image to sum to 1 and be within the range of 0 and 1. These outputs are then taken as analogous to trial probabilities and directly used as hit rates.

Similarity of two B.I1n vectors To measure the similarity of two B.I1n vectors, we use noise-corrected Spearman correlations. The same noise-correction formula is used for the RDM metric (Equation 4.2). The B.I1n metric is then defined as the correlation between a test feature space and the B.I1n vector estimated from monkeys performing 2-way alternative forced choice object recognition tasks on the same stimulus images.

Representational similarity analysis

Representational dissimilarity matrices (RDMs) represent the pair-wise dissimilarities (1 - Pearson r) between each pair of stimulus response vectors in an image-set [56, 159]. We measured the similarity between two RDMs (ex. a test brain area and a target brain area) as the Pearson correlation. To noise-correct this similarity measure, we estimate the self-consistency of our test and target RDMs by computing bootstrapped split-half-trial correlations (i.e., the correlation of two RDMs created using two disjoint sets of trials). With these estimates, we compute the noise corrected Pearson correlation between area A and area B's RDMs as,

$$r = \frac{r_{AB}}{\sqrt{r_{AA} r_{BB}}}, \tag{4.2}$$

where r_{AB} is the mean (across bootstraps) Pearson correlation between area A and B's RDMs. The same number of trials were used for area A in both the numerator and denominator, and similarly for area B. However, the number of trials for A and B were not necessarily the same (because we are correcting for differences in self-consistency, the trials do not need to be matched for this analysis).

In general, RDMs can summarize population responses to any arbitrary feature space, including brain areas and models. The same procedures apply to each (where the response vectors for models are vectors of feature responses instead of neural responses). For the V1-like control model we use in this study, there is no measurement noise and therefore its self-consistency is defined as 1.0.

We present data from the variation 640 dataset and evaluate all representations against the "cIT & aIT" pool using a bootstrapped sample of 265 units to estimate the target representation. We chose "cIT & aIT" as the target because most of our metrics are limited primarily by the number of neural sites and, for our data, we have the largest number of sites with this pool. As mentioned above, when computing the similarity of "cIT and aIT" against itself, we ensure that no data from the same piece of cortical tissue (implanted array) reside simultaneously in the r_{AA} and r_{BB} terms above, to help ensure all of our bootstrap samples are independent.

Single-site statistics

We performed two types of statistical tests on our single-site metrics (latency, sparsity, and d-prime measures): testing if the means are significantly different and testing if the right tails are significantly different. We use two-sided t-tests for the former and two-sample KS-tests for the latter.

For the t-tests, we first averaged the metric value for each neural site in each array implant together, resulting in a single mean metric value per array. We then used the array mean values as samples in our t-tests, so the number of samples for each area or age was simply the number of arrays we recorded. This test requires minimal assumption about the data, specifically not requiring us to assume each electrode within an array represents a statistically independent sample from the other electrodes within the array. In computing t-values, we did not assume equal variance between the test and target populations.

For the KS-tests, to fairly compare populations with differing numbers of neural sites, we sub-sampled populations when they were not matched. For instance, to compare area A to area B when A has more units than B, we randomly sampled neural sites from A to match the number of sites in B. After matching the areas in this way, we took the top 5% of units for a given metric and performed a two-sample KS-test with the two samples from A and B.

4.4 Results

To compare all of our neural data on a level playing field, all analyses utilize the bootstrap procedure described above to match the number of neural sites, number of trials (when the metric itself provides no measure of self-consistency for noise-correction), and number of stimuli used in each neural sample.

4.4.1 Single-site measures

We first present single-site visualizations and measures. Figure 4-2 shows the peri-stimulus time histograms from several neural sites and brain areas. Each area and channel exhibits heterogeneity across different views and positions of the same object, as indicated by the variable peaks of the individual lines in the plots. Qualitatively, the juvenile responses take longer to reach their peak values, when compared with any of the other adult areas.

We quantified the latency of neural responses using a threshold crossing metric. As shown in Figure 4-2, the juvenile pIT responses are more latent than any of the other adult areas (see Table 4.1 and 4.5 for statistical testing). As is visible in the figure, adult latencies for V4, pIT, and "cIT and aIT" almost always occur before 100

Table 4.1: Initial rise latency mean statistics

Test	Type of test	t-value	p-value
V4 vs. Juvenile pIT	Two-tailed t-test	-7.00	0.0134 *
V4 vs. Adult pIT	Two-tailed t-test	-2.95	0.1703
V4 vs. Adult aIT and cIT	Two-tailed t-test	-4.16	0.0089 **
Juvenile pIT vs. Adult pIT	Two-tailed t-test	3.54	0.0402 *
Juvenile pIT vs. Adult aIT and cIT	Two-tailed t-test	5.22	0.0228 *
Adult pIT vs. Adult aIT and cIT	Two-tailed t-test	0.93	0.4902

Table 4.2: Sparsity mean statistics

Test	Type of test	t-value	p-value
V4 vs. Juvenile pIT	Two-tailed t-test	16.83	0.0001 **
V4 vs. Adult pIT	Two-tailed t-test	0.45	0.7310
V4 vs. Adult aIT and cIT	Two-tailed t-test	8.72	0.0003 **
Juvenile pIT vs. Adult pIT	Two-tailed t-test	-1.87	0.3090
Juvenile pIT vs. Adult aIT and cIT	Two-tailed t-test	-7.73	0.0007 **
Adult pIT vs. Adult aIT and cIT	Two-tailed t-test	0.73	0.5963

Table 4.3: Object d-prime mean statistics

Test	Type of test	t-value	p-value
V4 vs. Juvenile pIT	Two-tailed t-test	0.30	0.7760
V4 vs. Adult pIT	Two-tailed t-test	-0.96	0.4228
V4 vs. Adult aIT and cIT	Two-tailed t-test	-1.52	0.1889
Juvenile pIT vs. Adult pIT	Two-tailed t-test	-1.26	0.3162
Juvenile pIT vs. Adult aIT and cIT	Two-tailed t-test	-1.84	0.1255
Adult pIT vs. Adult aIT and cIT	Two-tailed t-test	-0.54	0.6287

Table 4.4: Category d-prime mean statistics

Test	Type of test	t-value	p-value
V4 vs. Juvenile pIT	Two-tailed t-test	0.77	0.4844
V4 vs. Adult pIT	Two-tailed t-test	-1.58	0.2260
V4 vs. Adult aIT and cIT	Two-tailed t-test	-1.41	0.2202
Juvenile pIT vs. Adult pIT	Two-tailed t-test	-2.87	0.0702
Juvenile pIT vs. Adult aIT and cIT	Two-tailed t-test	-2.03	0.1058
Adult pIT vs. Adult aIT and cIT	Two-tailed t-test	-0.58	0.6003

Table 4.5: Initial rise latency right tail statistics

Test	Type of test	D-value	p-value
V4 vs. Juvenile pIT	Two-sample KS-test	0.86	0.0042 **
V4 vs. Adult pIT	Two-sample KS-test	0.86	0.0042 **
V4 vs. Adult aIT and cIT	Two-sample KS-test	0.83	0.0122 *
Juvenile pIT vs. Adult pIT	Two-sample KS-test	1.00	0.0004 **
Juvenile pIT vs. Adult aIT and cIT	Two-sample KS-test	1.00	0.0013 **
Adult pIT vs. Adult aIT and cIT	Two-sample KS-test	0.83	0.0122 *

Table 4.6: Sparsity right tail statistics

Test	Type of test	D-value	p-value
V4 vs. Juvenile pIT	Two-sample KS-test	1.00	0.0001 **
V4 vs. Adult pIT	Two-sample KS-test	0.71	0.0275 *
V4 vs. Adult aIT and cIT	Two-sample KS-test	0.75	0.0098 **
Juvenile pIT vs. Adult pIT	Two-sample KS-test	1.00	0.0004 **
Juvenile pIT vs. Adult aIT and cIT	Two-sample KS-test	1.00	0.0002 **
Adult pIT vs. Adult aIT and cIT	Two-sample KS-test	0.86	0.0042 **

Table 4.7: Object d-prime right tail statistics

Test	Type of test	D-value	p-value
V4 vs. Juvenile pIT	Two-sample KS-test	0.44	0.2500
V4 vs. Adult pIT	Two-sample KS-test	1.00	0.0004 **
V4 vs. Adult aIT and cIT	Two-sample KS-test	1.00	0.0002 **
Juvenile pIT vs. Adult pIT	Two-sample KS-test	0.57	0.1287
Juvenile pIT vs. Adult aIT and cIT	Two-sample KS-test	0.75	0.0098 **
Adult pIT vs. Adult aIT and cIT	Two-sample KS-test	0.43	0.4232

Table 4.8: Category d-prime right tail statistics

Test	Type of test	D-value	p-value
V4 vs. Juvenile pIT	Two-sample KS-test	0.44	0.2500
V4 vs. Adult pIT	Two-sample KS-test	0.71	0.0275 *
V4 vs. Adult aIT and cIT	Two-sample KS-test	1.00	0.0002 **
Juvenile pIT vs. Adult pIT	Two-sample KS-test	0.43	0.4232
Juvenile pIT vs. Adult aIT and cIT	Two-sample KS-test	0.75	0.0098 **
Adult pIT vs. Adult aIT and cIT	Two-sample KS-test	0.43	0.4232

ms post-stimulus onset. In contrast, juvenile pIT latencies are almost entirely after 100 ms post-stimulus onset.

We present also sparsity measures in Figure 4-2 and significance testing in Tables 4.2 and 4.6. While the mean sparsity significantly differed, in some cases, between adult V4 and adult IT, we found no such difference between adult and juvenile pIT. However, we did find a significant difference between the right tails of the juvenile and adult pIT distributions. With the right tail statistics, we found significant differences between all 6 possible area comparisons, perhaps indicating that sparsity, in fact, differs at each area, or alternatively that the measure is sensitive to properties shared across electrodes within each array.

To assess single-site object and category selectivity, we measured object and category d-primes (Figure 4-2). These bootstrapped measures quantify the extent a neural site discriminates between its top responding object or category and the remaining objects or categories. We used separate trials to determine each site's top responding object or category and measured d-primes on the remaining trials. Unlike latency, the mean d-primes did not show a significant difference between the juvenile and adult data (Tables 4.3 and 4.4). However, the right tail statistics did differ, in some cases, between adult IT and V4 (Tables 4.7 and 4.8). No significant differences were found between adult and juvenile pIT.

4.4.2 Performance

Performance measures provide us a means to directly query a neural sample's capacity to support behavioral recognition tasks—the hypothesized function of higher visual cortex. We measured the accuracy of linear classifiers in an 8-way categorization task, as shown in Figure 4-3. Our results do not indicate a significant difference between juvenile and adult pIT samples when using responses taken from 70-170 ms post-stimulus onset.

In light of our prior latency results, we measured performance using responses in three additional time windows: 50-100 ms, 100-150 ms, and 150-200 ms. We observe that only the first window, 50-100 ms shows a difference in performance readout

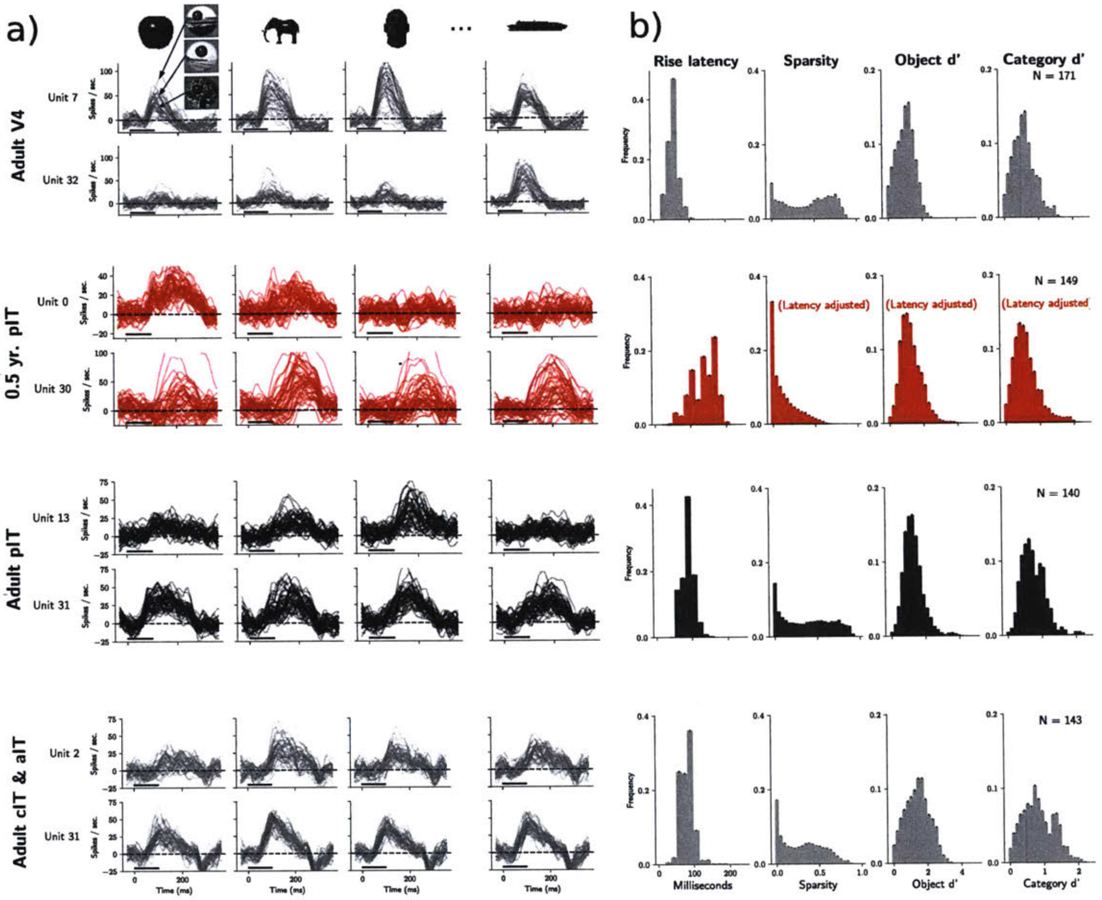

Figure 4-2: Single-site measures

a. Shown are the peri-stimulus time histograms (PSTHs). The rows of the PSTH plots represent responses from the same neural site. The columns represent responses to all images depicting the same object (shown at the top of each column–each object is shown in 40 distinct images). Each line in the PSTH plots represents the mean (across trial) response to a single stimulus image. The black bar at the bottom left of each PSTH plot indicates the time at which the stimulus was shown (100 ms, followed by 100 ms of gray screen). **b.** Shown are onset latency, sparsity, object d-prime, and category d-prime estimates across bootstrapped samples of trials. Each column indicates one metric (latency, sparsity, object, or category d-prime) and each row indicates an area or age-defined pool of neural sites (ex. V4 or juvenile pIT). Plots indicate the frequency distribution across units with error bars indicating the standard error of the mean over bootstrapped draws of random trials.

between juvenile and adult pIT.

4.4.3 Behavioral similarity

We used a behavioral similarity measure to assess if decoders trained on neural samples made similar errors to monkeys performing an object recognition task (see Figure 4-4). Similar to our performance measures, there is no significant difference between our juvenile and adult pIT samples.

While, in some ways, our behavioral similarity measure offers more granularity than the performance measure presented above (representations with equally high performance would not be dissociable by a performance measure, but could be dissociable by a measure which takes into account the pattern of errors as our behavioral measure does here), this behavioral measure is somewhat deficit in terms of power. For instance, it is not able to separate IT from V4 as robustly as a performance-only metric (compare Figures 4-3 and 4-4). The power difference between the two metrics might be due to measurement error existing in both the test (neural samples) and target (behavioral confusions), as compared with categorization performance where no measurement error exists in the target (category labels).

4.4.4 Representational dissimilarity

Representational dissimilarity matrices (RDMs) are a compact way to summarize feature spaces. The technique, termed representational similarity analysis (RSA) uses RDMs to quantify similarity between brain areas and models (or between brain areas and brain areas) [56, 159]. Similar to the other population-level metrics presented above (performance and behavioral consistency), we do not observe significant differences between our juvenile and adult pIT samples (see Figure 4-5). Because we have the highest number of neural samples in the combined areas of adult cIT and aIT, we designated this area as the target to which we measure similarities against. Consequently, the similarity between the "cIT & aIT" pool and itself (see the bottom left plot of Figure 4-5) can be interpreted as the noise-ceiling for this metric. As

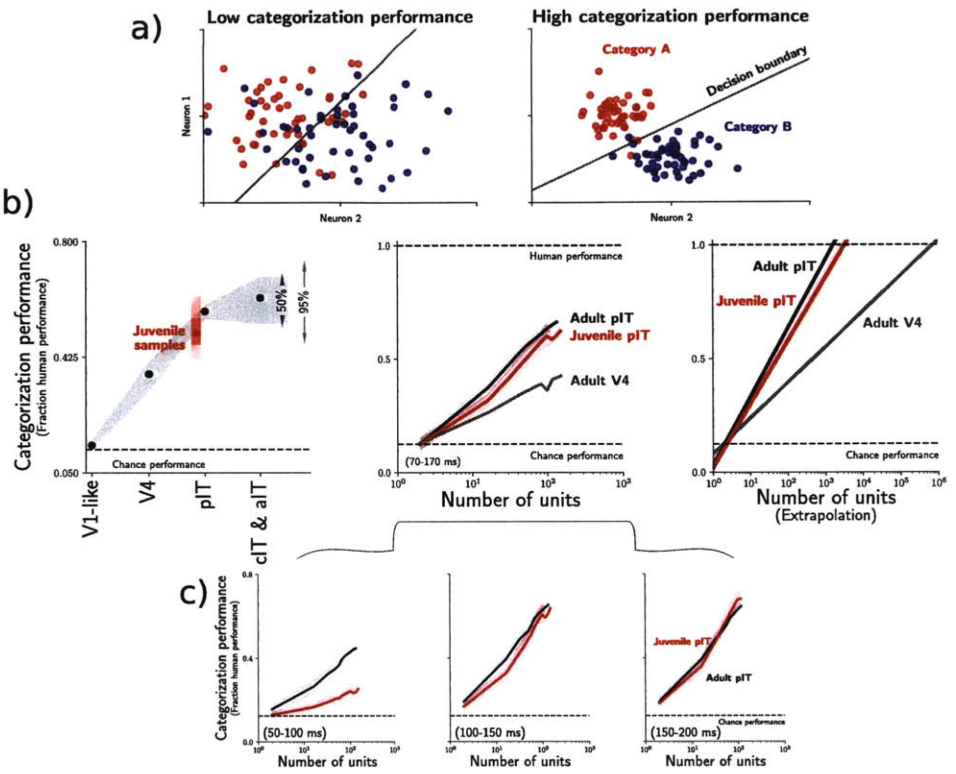

Figure 4-3: Categorization performance

a. Conceptual, simulated example of linear classification. Each plot shows the stimulus responses for a pair of two neurons (x- and y-axes). Dots indicate the response to a particular stimulus image and the color (red or blue) indicates the category which the image depicts. Shown also is the line which best separates the two categories. The plot on the left illustrates a pair of neurons which are not very categorical, in the sense that a linear classifier can not easily separate the categories, in contrast with the plot on the top right. This concept of linear separability generalizes to additional neurons (higher dimensions) which we utilize for our population-level classifications. **b. left.** Performance on 8 category discrimination tasks using the variation 3 stimulus set. Shown are the cross-validated performances of each area normalized to human performance. The light gray shading indicates the spread of 95% of bootstrapped performance values, while the dark gray indicates the spread of 50% of performance values for each area. To control for experimental noise and sampling, all areas were matched to contain the same number of trials and units. **b. center.** Performance as a function of the number of neural sites used. The width of the shading around each line indicates the spread of 50% of bootstrapped performance values. All areas were matched to contain the same number of trials. **b. right.** Extrapolated performance for each area. The extrapolation indicates log-linear fits to the plot in the center column. **c.** Shown are categorization performances as a function of units using different read-out windows of time (50-100 ms, 100-150 ms, and 150-200 ms).

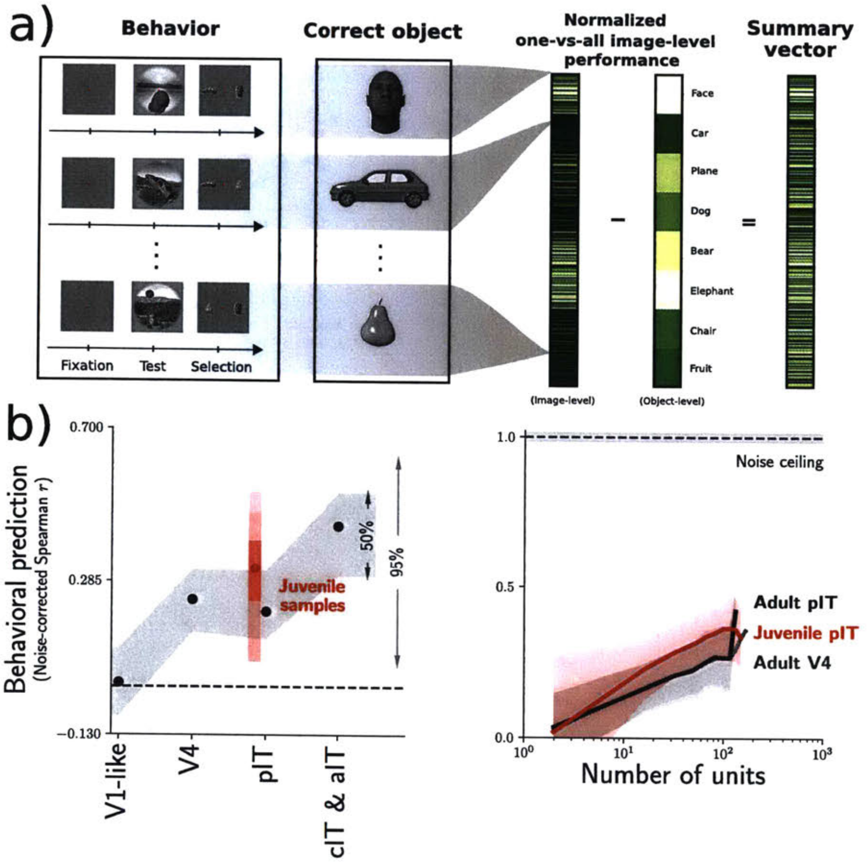

Figure 4-4: Behavioral consistency

a. Creation of a B.I1n behavioral response vector. Animals initiate a 2-way alternative fixed choice (2AFC) trial by fixating at a red dot at the center of the screen. A test image is then shown for 100 ms followed by a selection phase. The selection phase presents two objects in a canonical position on a gray background. One object was previously shown in the test image, and a second distractor object is also shown. The animal indicates its choice as to which object was shown in the test image by saccading to the left or right. We then construct a vector with each element representing the performance of recognizing the object in a particular test image (see Methods). This vector is then normalized by the average performance (across all images depicting the object) to normalize out object-level differences and focus the vector's variance more so on image-level variance. **b. left.** Shown are the noise-corrected behavioral similarities of each area to monkey behavioral responses on variation 640 recognition tasks. The light gray shading indicates the spread of 95% of bootstrapped correlation values, while the dark gray indicates the spread of 50% of correlation values for each area. To control for experimental noise and sampling, all areas were matched to contain the same number of trials and units. **b. right.** Behavioral similarity as a function of the number of neural sites used. The width of the shading around each line indicates the spread of 50% of bootstrapped correlation values. All areas were matched to contain the same number of trials.

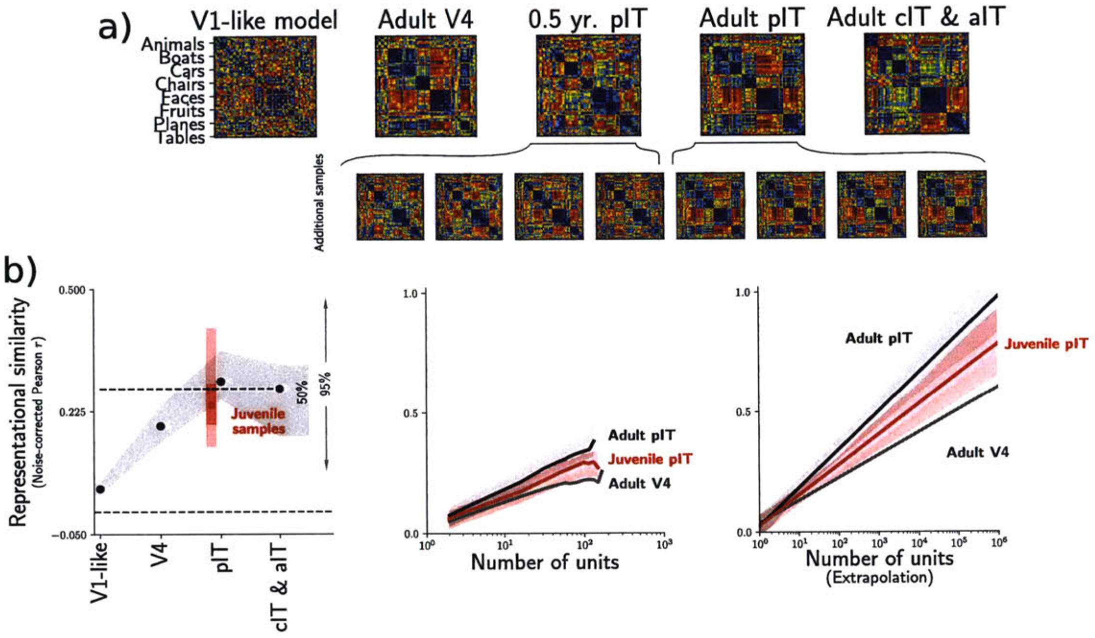

Figure 4-5: Representational similarity

a. Object-averaged RDMs for each brain or age-defined area. Each RDM was created from samples of 37 trials and 58 units on the variation 3 dataset. Also shown below are RDMs created from additional random samples using the same number of units and trials. **b. left.** Shown are the noise-corrected RDM correlations of each area to the pool of adult cIT and aIT neural sites on variation 640 (the stimulus set over which we have recorded the most neural sites). We chose to use the combined data from adult cIT and adult aIT as the target RDM because this allowed us to utilize the most data (265 target neural sites). The light gray shading indicates the spread of 95% of bootstrapped correlation values, while the dark gray indicates the spread of 50% of correlation values for each area. To control for experimental noise and sampling, all areas were matched to contain the same number of trials and units. **b. center.** RDM similarity as a function of the number of neural sites used. The width of the shading around each line indicates the spread of 50% of bootstrapped correlation values. **b. right.** Extrapolated correlations for each area. The extrapolation indicates log-linear fits to the plot in the center column.

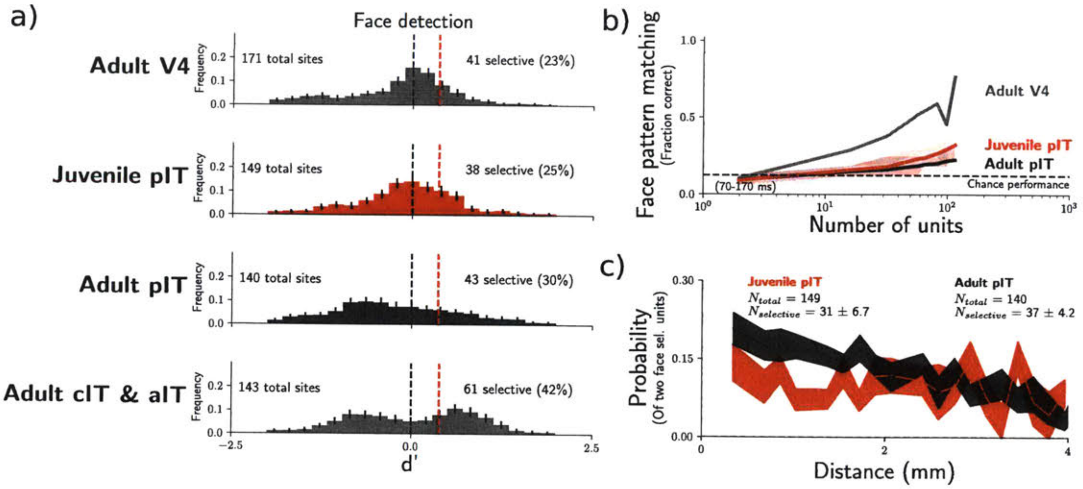

Figure 4-6: Face detection and identification

a. Shown is the distribution over neural sites of face detection d-primes. The face detection d-prime is the d-prime between face and non-face stimuli. All plots were repetition matched to use the same number of trials. The red dashed line indicates the 99th percentile of shuffled d-primes where we computed face-vs.-non-face d-primes on data which had the stimulus identities randomly scrambled. The number of selective units was defined as the number of units above the 99th percentile of shuffled d-primes. Similar to the object and category d-primes presented above, we latency adjusted each neural site before computing d-prime estimates—this correction was not used for the identification estimations in panel b. **b.** Shown are the face identification performances of each area and age-defined pool, estimated from linear classifiers. In the variation 0 dataset, there were 8 face identities. **c.** Shown is the spatial clustering of face selective units within array implants. Starting from each face selective site, we plot the probability of finding another face selective site as a function of spatial distance.

shown in the bottom left plot of Figure 4-5, the median of our juvenile pIT samples is within the noise ceiling of the cIT & aIT pool (it is within the spread containing 95% of the bootstrapped mass of the ceiling).

4.4.5 Face identification and detection

To compare with prior literature on face selectivity, we used our variation 0 dataset to compute face detection d-primes and face identification performance. Variation 0 depicts all objects in a canonical, centrally positioned, forward facing view, similar to the presentation of faces in prior studies. We found that juvenile face detection d-primes (see Figure 4-6) fell within the distribution of adult data. In general, we found that face d-prime measures to be poor discriminators between even adult areas (compare adult V4 and "cIT & aIT" where we would expect the largest differences), consistent with what we found with other single-site d-prime measures (see Figure 4-2).

An additional measure of face recognition performance is face identification: recognizing the identity of a specific face. To estimate face identification performance, we trained linear classifiers using similar methodology to our categorization classifiers described above. We found that adult V4 excelled most at this task, consistent with prior work which has shown V4 outperforming IT in low variation subordinate recognition tasks [9]. Juvenile pIT was no more or less performant than adult pIT on this recognition task (see Figure 4-6), consistent with our prior categorization results described above.

To assess the spatial distribution of face selective sites, we plotted the probability of finding two units as a function of distance. Representations in which face selective units cluster spatially together (as is known to occur in adults), would exhibit a negative slope, that is, the probability of finding another face selective site goes down the farther away one moves from a face selective site. At the other extreme, representations in which units are distributed randomly would exhibit a flat line—there would be equal probability of finding another face selective site regardless of where one searches. Our results are suggestive of a difference in spatial organization between adults and juveniles. However, we caution that because we restricted our analyses to comparisons within single array implants (which have grids of 10 by 10 electrodes) and that out of this grid, not all sites contained visually responding units (with even fewer being face selective), our power to make claims relating to spatial layouts is limited.

4.5 Discussion

In this study, we have provided the first large-scale, detailed account of juvenile pIT on a challenging image-set capable of assaying the limits of core object recognition. We focused on population-level rate coding metrics to quantify consistency between juvenile neural responses, adults neural responses, and behavior (with the link between the latter two previously established [8]). We additionally compared our juvenile to adult data using single-site measures, primarily to connect with earlier work. We have presented evidence that, at the age of 19-32 weeks, pIT supports robust categorization read-outs in a manner presently indistinguishable from adults. We observe latency effects consistent with prior results [168], while providing additional evidence that IT exhibits more latent responses at the age of 19-32 weeks. [168] pooled data from infants recorded in the range of 4-28 weeks, with most of the recordings occurring before the age of 16 weeks, leaving open the possibility that the adult-infant latency gap may be less pronounced at the ages we recorded from. We note that the slower juvenile onset latencies do result in decreased recognition (relative to adults) 50-100 ms after stimulus onset (see Figure 4-3), they do not cause decrements in performance in later time bins (ex. 100-150 ms). The behavioral significance, or lack thereof, of this latency effect will require future investigation.

In terms of future developmental electrophysiological studies, there are at least three possible directions of interest: (1) record from earlier ages in hopes that developmental effects will be more easily observable at earlier ages, (2) record more adult and juvenile data at the same ages in hopes that there may be differences in representation at this age not presently detectable due to experimental noise, (3) use alternative experimental techniques (such as an exchangeable array [182] or large-scale single-site approaches similar to [183]) to increase not only the number of sampled neural sites but also to help ensure that each neural sample is closer to the idealized independent, "random" sample from pIT. Point (3) may also allow for a more detailed characterization of both the adult and juvenile states in terms of measures not investigated here: such as cell type or layer differences.

4.5.1 Response onset latency

Both the cause and significance, if any, of neuronal latency to development remain unknown. Possible causes may include differences in synaptic efficacy [184, 185, 186], myelination diameter, internodal distance [187], and total axon diameter — all factors known to change with development. A combination of these factors may be responsible for developmental latency [187].

4.5.2 Single-site metrics

Aside from neuronal response latencies, for which we see robust differences between juveniles and adults, we did not find differences in the means of any of our d-prime measures (Tables 4.3 and 4.4). We additionally did not see reliable differences between adult areas using the mean values of these metrics—even between V4 and "cIT & aIT" for which we would expect to see the largest of differences among our adult data. While there are many examples in the literature of using single-site measures to characterize both low [126, 188, 63, 64] and high-level visual areas [189, 30, 5, 168], these approaches do not appear to provide strong diagnostic utility here (even between adult areas) unless statistics are computed on the, inherently noisier, right tails of the distributions (Tables 4.7 and 4.8).

We similarly did not find significant differences between the mean sparsity of juvenile and adult pIT (Table 4.2), but found significant differences when computing statistics on the right tails (Table 4.6). If these differences are reflective of actual biological differences between both areas and ages, rather than being a reflection of individual differences in character or quality of each array implant (for which we tried to equalize using trial matching and unified electrode selection procedures), then these results could connect to prior work on sparse coding. Prior modeling work has demonstrated that V1-like edge-selective tuning properties arise, in part, from response sparsity constraints [45, 190] and have explored its utility as a regularizer in machine learning [191]. However, the connection, if any to biological development remains to be established and requires further examination in narrowing down these and other possibilities.

4.5.3 Face metrics

While single-site metrics do not provide us the ability to separate V4 from IT here, these measures do indicate that all areas measured contain face discrimination and identification beyond what would be expected from chance (see Figure 4-6). These results do not contradict prior work, finding 1 year old macaques exhibiting face selective fMRI regions [42].

Human infant fMRI at 12-32 weeks of age (corresponding to approximately 3-8 weeks in age for macaques), also exhibit face selective fMRI regions [37]. However, [37] found that infants had less selective responses which additionally corresponded to differing category-level representational similarities between adults and infants. It is an open question as to whether or not electrophysiological recordings in 3-8 week old macaques would also yield such differences. There are are at least two scenarios in which we would expect to observe deficits in infant fMRI face patch measurements while simultaneously observing adult-like physiological responses: (1) fMRI deficits may reflect changes in vascular development unrelated to the visual representation itself, (2) fMRI may be detecting large-scale "organizational" differences between adult and infants, with the underlying neural selectivities of infants and adults being indistinguishable. Reorganization of this nature (whereby spatially nearby neurons represent similar visual features) may have biological utility in minimizing metabolic activity but have no relation to recognition performance. Many high-performing object recognition models contain one or more fully-connected layers as the last layers of the model where the spatial layout of the representation is discarded entirely [6]—providing evidence that, computationally, spatial organization (in higher layers) is not necessary for robust object recognition performance.

While our experimental approaches offer some spatial information allowing us to test the second hypothesis above (Figure 4-6c), our power is limited because we are restricted in making cross-animal or hemisphere comparisons. With this preface, we do observe that our juvenile recordings have reduced spatial clustering of face selective units. This finding is not inconsistent with the second hypothesis described in the paragraph above and, if generally reflective of juvenile pIT, may explain some of the discrepancies suggested by physiological and fMRI studies.

4.5.4 Constraints on learning algorithms

While progress has been made in understanding and modeling both higher [1] and lower visual areas [4, 5, 2, 3, 10, 8] aided, in part, by advances in computer vision [6, 7], current high-performing (and neurally predictive) models require implausibly high amounts of precisely labeled training examples. Thus, visual developmental research may serve to both constrain and guide the development of more generic and predictive learning algorithms of primate vision, and possibly beyond.

Provided that we did not see differences using population-level metrics, our work does not constrain models of development in a representational sense. However, our work does constrain models from the perspective of the amount of training data—models of primate visual development ought to exhibit adult-like performance and consistency using only 25 weeks of visual experience. Assuming 16 hours of awake visual experience per day [192] over 25 weeks, our data constrains models of primate visual development to roughly 2,800 hours of waking visual experience.

Chapter 5

Conclusions

This thesis has examined several aspects of both adult (Chapter 2) and juvenile (Chapter 4) visual object recognition. We have additionally described (Chapter 3) a set of natural image and input statistics automatically incorporated into high-performing models through supervised training. Our adult fMRI studies have extended prior work [10] which showed increased model recognition performance (across model architectures) was a predictor of increased neural predictivity of IT and V4. We furthered this line of investigation by demonstrating that neural predictivity also increases within the same model architecture as it is trained to perform object recognition tasks on natural images. Our work on natural image and input statistics demonstrated that some, but not all, of model performance was sustained only by matching 2^{nd} order filter statistics. This work might lend some plausibility to visual object recognition being learnable from partly generic learning rules, as opposed to being largely or entirely dictated by evolutionary architectural constraints.

Our juvenile recordings in higher visual cortex provide the first detailed account of the juvenile visual object representation at the scale of multi-unit recordings. With our finding that the multi-unit population representation is indistinguishable from the adult representation, we are able to rule out several broad-scale hypotheses predicting the time-course of visual development (see Figure 5-1). Given that we did not find a representational difference between juveniles and adults, the possibility of a complete "innatism" model of visual object recognition development (adult-like responses at

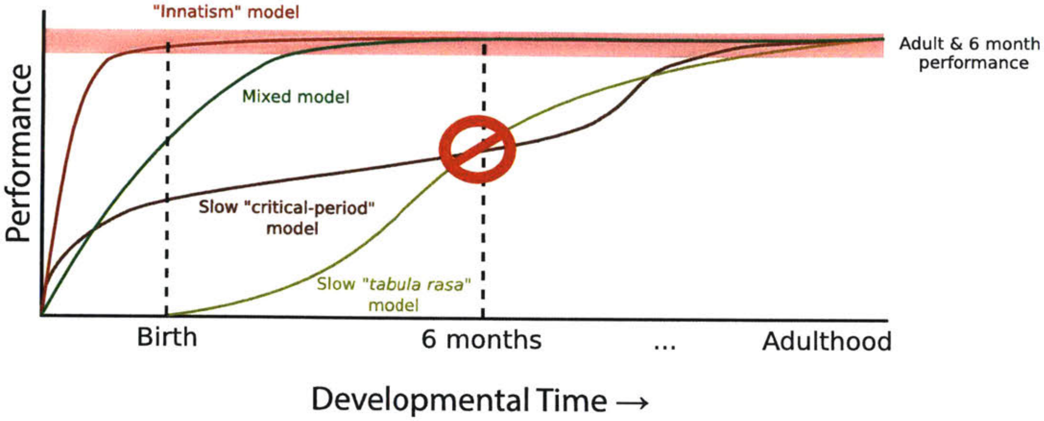

Figure 5-1: Constrained neural developmental time-courses
Our juvenile pIT recordings rule out hypotheses predicting sub-adult recognition performances at 6 months of age. However, the data do no distinguish between any hypotheses which predict adult-like performance at 6 months.

birth) cannot still be ruled out. While studies of early visual areas find refinement of, for example, orientation tuning in the first few weeks of postnatal life, it is possible that these changes may not be necessary for invariant object recognition performance and that sharpening of early visual area tuning curves could be a consequence of entirely unrelated processes (such as optimizing metabolic processes or in refining synaptic efficiency). It may also be possible that the refinement tuning in early visual areas does serve some behavioral function, but that our tasks were not sufficient to probe these functions (perhaps the refinement is beneficial for either more difficult recognition tasks, or for visual tasks unrelated to object recognition entirely, such as motion detection).

There are multiple ways in which visual development may be further investigated in light of our results, some of which are described in Chapter 4. While we have ruled out several "slow" models of visual development (see Figure 5-1), similar variations of these are still plausible, with an accelerated temporal time-course (see Figure 5-2). These "faster" hypotheses could be tested by replicating our methods in animals earlier in age. Aside from recording from younger animals, an alternative approach

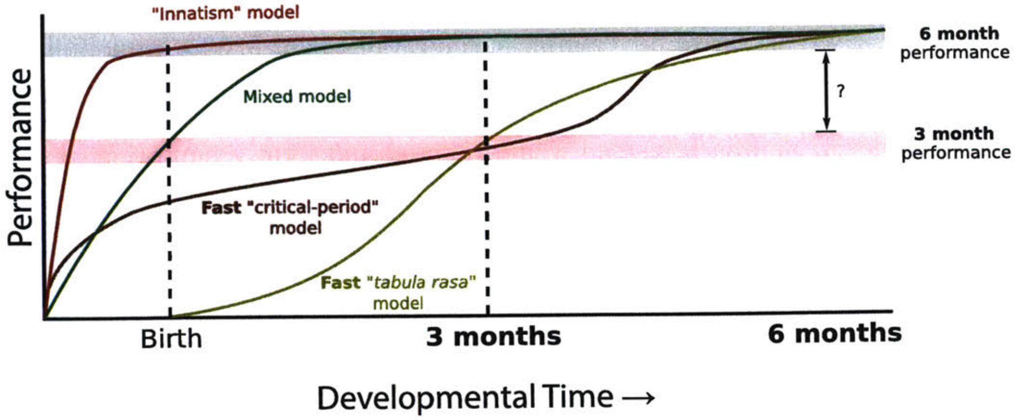

Figure 5-2: Potential earlier neural developmental time-courses
With the constraints of our data, similar hypotheses can be proposed and tested by recording from animals younger in age (for example, 3 months), similar to our initial set of possibilities shown in Figure 1-2.

would be to record from additional animals at the same ages we have collected here—in the context of Figure 5-1, this would result in us reducing the vertical width of the red and black bars (presently laying exactly on top of each other in the figure). Conceivably, a gap could emerge between the representations with additional data.

With the prior discussion in mind, it is somewhat straightforward to further constrain models of visual development, however, it is less clear as to how to actually build more predictive models. Supposing there are, in fact, behaviorally-relevant developmental changes to higher visual cortex in postnatal life, two broad approaches can be taken: a bottom up approach and a top down approach to studying visual development (in practice, most approaches will necessarily have elements of each and the distinction is somewhat subjective).

The "bottom up" approach mentioned here is used to refer to a broader class of approaches than is typically meant by the term. Here, "bottom up" refers to approaches that focus on optimizing models on loss functions of explicit visual nature. The loss functions could either be semi-supervised or completely unsupervised. Examples include optimizations to explicitly incorporate natural image statistics into models (like Chapter 3), optimizations which seek to model or predict visual inputs through time [49], generative adversarial optimizations [52], pixel completion tasks [50], among other approaches.

The "top down" approach instead would proceed under the assumption, or hope, that visual functions would emerge from loss functions which are not explicitly visual. For example, from a reinforcement optimization with rewards being specified by both external and internal rewards (the latter of which might be largely specified by evolution and could include targets like curiosity [193], or other forms of intrinsic motivation [194]). It is important to note that reinforcement learning in itself does not distinguish the top down and bottom up approaches here. For example, "supervised" category classification could be distilled down to a reinforcement problem where all the labels are provided in the form of a reward signal indicating a correct or incorrect response. Rather, what distinguishes the approaches is the (somewhat subjective) generality of the learning signal. The example of giving rewards for category classification, or any other explicitly visual task, is an example of a lack of this generality.

The top down approach has the subjective appeal of potentially demonstrating much or all of neural visual computation arises from more generic survival or adaptation constraints. In the same way that [10] and Chapter 2 suggest that middle and early visual areas can be understood as subserving a higher-level objective (visual object recognition), object recognition itself may be demonstrated to be an epiphenomenon of a more generic, cross-modality optimization.

Bibliography

[1] M. Carandini, J. B. Demb, V. Mante, D. J. Tolhurst, Y. Dan, B. A. Olshausen, J. L. Gallant, and N. C. Rust, "Do we know what the early visual system does?," *The Journal of Neuroscience*, vol. 25, no. 46, pp. 10577–10597, 2005.

[2] J. J. DiCarlo, D. Zoccolan, and N. C. Rust, "How does the brain solve visual object recognition?," *Neuron*, vol. 73, no. 3, pp. 415–434, 2012.

[3] T. O. Sharpee, M. Kouh, and J. H. Reynolds, "Trade-off between curvature tuning and position invariance in visual area V4," *Proceedings of the National Academy of Sciences*, no. 28, pp. 11618–11623, 2013.

[4] C. P. Hung, G. Kreiman, T. Poggio, and J. J. DiCarlo, "Fast readout of object identity from macaque inferior temporal cortex," *Science*, vol. 310, no. 5749, pp. 863–866, 2005.

[5] N. C. Rust and J. J. DiCarlo, "Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area V4 to IT," *The Journal of Neuroscience*, vol. 30, no. 39, pp. 12978–95, 2010.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.

[7] J. Bergstra, D. Yamins, and D. D. Cox, "Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures," in *JMLR*, 2013.

[8] N. J. Majaj, H. Hong, E. A. Solomon, and J. J. DiCarlo, "Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance," *The Journal of Neuroscience*, vol. 35, no. 39, pp. 13402–13418, 2015.

[9] H. Hong, D. L. Yamins, N. J. Majaj, and J. J. DiCarlo, "Explicit information for category-orthogonal object properties increases along the ventral stream," *Nature Neuroscience*, vol. 19, no. 4, p. 613, 2016.

[10] D. L. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo, "Performance-optimized hierarchical models predict neural responses in higher visual cortex," *Proceedings of the National Academy of Sciences*, vol. 111, no. 23, pp. 8619–8624, 2014.

[11] C. F. Cadieu, H. Hong, D. L. Yamins, N. Pinto, N. J. Majaj, and J. J. DiCarlo, "The Neural Representation Benchmark and its Evaluation on Brain and Machine," *International Conference on Learning Representations (ICLR)*, vol. 65, no. 6, 2013.

[12] G. Golarai, D. G. Ghahremani, S. Whitfield-Gabrieli, A. Reiss, J. L. Eberhardt, J. D. Gabrieli, and K. Grill-Spector, "Differential development of high-level visual cortex correlates with category-specific recognition memory," *Nature Neuroscience*, vol. 10, no. 4, p. 512, 2007.

[13] Z. Kourtzi, M. Augath, N. K. Logothetis, J. A. Movshon, and L. Kiorpes, "Development of visually evoked cortical activity in infant macaque monkeys studied longitudinally with fMRI," *Magnetic Resonance Imaging*, vol. 24, no. 4, pp. 359–366, 2006.

[14] N. Li and J. J. DiCarlo, "Unsupervised natural visual experience rapidly reshapes size-invariant object representation in inferior temporal cortex," *Neuron*, vol. 67, no. 6, pp. 1062–1075, 2010.

[15] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *2011 IEEE International Conference on Computer Vision (ICCV)*, pp. 2018–2025, IEEE, 2011.

[16] P. Mazzoni, R. A. Andersen, and M. I. Jordan, "A more biologically plausible learning rule for neural networks.," *Proceedings of the National Academy of Sciences*, vol. 88, no. 10, pp. 4433–4437, 1991.

[17] F. Anselmi, J. Z. Leibo, L. Rosasco, J. Mutch, A. Tacchetti, and T. Poggio, "Unsupervised learning of invariant representations," *Theoretical Computer Science*, vol. 633, pp. 112–121, 2016.

[18] A. Coates, A. Karpathy, and A. Y. Ng, "Emergence of object-selective features in unsupervised feature learning," in *Advances in Neural Information Processing Systems*, pp. 2681–2689, 2012.

[19] D. D. Cox, P. Meier, N. Oertelt, and J. J. DiCarlo, "'Breaking' position-invariant object recognition," *Nature Neuroscience*, vol. 8, no. 9, p. 1145, 2005.

[20] Y. Yan, M. J. Rasch, M. Chen, X. Xiang, M. Huang, S. Wu, and W. Li, "Perceptual training continuously refines neuronal population codes in primary visual cortex," *Nature Neuroscience*, vol. 17, no. 10, p. 1380, 2014.

[21] N. Li and J. J. DiCarlo, "Unsupervised natural experience rapidly alters invariant object representation in visual cortex," *Science*, vol. 321, no. 5895, pp. 1502–1507, 2008.

[22] L. Wiskott and T. J. Sejnowski, "Slow feature analysis: Unsupervised learning of invariances," *Neural Computation*, vol. 14, no. 4, pp. 715–770, 2002.

[23] M. C. Crair, D. C. Gillespie, and M. P. Stryker, "The role of visual experience in the development of columns in cat visual cortex," *Science*, vol. 279, no. 5350, pp. 566–570, 1998.

[24] D. L. Ringach, "You get what you get and you don't get upset," *Nature Neuroscience*, vol. 14, no. 2, p. 123, 2011.

[25] A. K. Kreile, T. Bonhoeffer, and M. Hübener, "Altered visual experience induces instructive changes of orientation preference in mouse visual cortex," *The Journal of Neuroscience*, vol. 31, no. 39, pp. 13911–13920, 2011.

[26] C. J. Shatz and M. P. Stryker, "Ocular dominance in layer IV of the cat's visual cortex and the effects of monocular deprivation.," *The Journal of Physiology*, vol. 281, no. 1, pp. 267–283, 1978.

[27] S. Le Vay, T. N. Wiesel, and D. H. Hubel, "The development of ocular dominance columns in normal and visually deprived monkeys," *Journal of Comparative Neurology*, vol. 191, no. 1, pp. 1–51, 1980.

[28] L. Kiorpes, "The puzzle of visual development: behavior and neural limits," *Journal of Neuroscience*, vol. 36, no. 45, pp. 11384–11393, 2016.

[29] H. R. Rodman and M. J. Consuelos, "Cortical projections to anterior inferior temporal cortex in infant macaque monkeys," *Visual Neuroscience*, vol. 11, no. 1, pp. 119–133, 1994.

[30] H. R. Rodman, J. P. Skelly, and C. G. Gross, "Stimulus selectivity and state dependence of activity in inferior temporal cortex of infant monkeys," *Proceedings of the National Academy of Sciences*, vol. 88, no. 17, pp. 7572–7575, 1991.

[31] J. H. Marshel, M. E. Garrett, I. Nauhaus, and E. M. Callaway, "Functional specialization of seven mouse visual cortical areas," *Neuron*, vol. 72, no. 6, pp. 1040–1054, 2011.

[32] A. B. Saleem, A. Ayaz, K. J. Jeffery, K. D. Harris, and M. Carandini, "Integration of visual motion and locomotion in mouse visual cortex," *Nature Neuroscience*, vol. 16, no. 12, p. 1864, 2013.

[33] G. T. Prusky, P. W. West, and R. M. Douglas, "Behavioral assessment of visual acuity in mice and rats," *Vision Research*, vol. 40, no. 16, pp. 2201–2209, 2000.

[34] K. Ohki, S. Chung, Y. H. Ch'ng, P. Kara, and R. C. Reid, "Functional imaging with cellular resolution reveals precise micro-architecture in visual cortex," *Nature*, vol. 433, no. 7026, p. 597, 2005.

[35] H. Ko, S. B. Hofer, B. Pichler, K. A. Buchanan, P. J. Sjöström, and T. D. Mrsic-Flogel, "Functional specificity of local synaptic connections in neocortical networks," *Nature*, vol. 473, no. 7345, p. 87, 2011.

[36] N. Pinto, D. D. Cox, and J. J. DiCarlo, "Why is real-world visual object recognition hard?," *PLoS computational biology*, vol. 4, no. 1, p. e27, 2008.

[37] B. Deen, H. Richardson, D. D. Dilks, A. Takahashi, B. Keil, L. L. Wald, N. Kanwisher, and R. Saxe, "Organization of high-level visual cortex in human infants," *Nature Communications*, vol. 8, p. 13995, 2017.

[38] M. Livingstone, J. Vincent, T. Savage, and K. Srihasam, "Development of category-selective domains in infant macaque inferotemporal cortex," *Journal of Vision*, vol. 14, no. 10, pp. 228–228, 2014.

[39] M. J. Arcaro, P. F. Schade, J. L. Vincent, C. R. Ponce, and M. S. Livingstone, "Seeing faces is necessary for face-domain formation," *Nature Neuroscience*, vol. 20, no. 10, p. 1404, 2017.

[40] L. Kiorpes, A. Movshon, and W. Chaluppa, "Neuronal limitations on visual development in primates: Beyond striate cortex," 2014.

[41] J. S. Espinosa and M. P. Stryker, "Development and plasticity of the primary visual cortex," *Neuron*, vol. 75, no. 2, pp. 230–249, 2012.

[42] K. Srihasam, J. B. Mandeville, I. A. Morocz, K. J. Sullivan, and M. S. Livingstone, "Behavioral and anatomical consequences of early versus late symbol training in macaques," *Neuron*, vol. 73, no. 3, pp. 608–619, 2012.

[43] P. Agrawal, R. Girshick, and J. Malik, "Analyzing the performance of multilayer neural networks for object recognition," in *European conference on computer vision*, pp. 329–344, Springer, 2014.

[44] W. E. Vinje and J. L. Gallant, "Sparse coding and decorrelation in primary visual cortex during natural vision," *Science*, vol. 287, no. 5456, pp. 1273–1276, 2008.

[45] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?," *Vision research*, vol. 37, no. 23, pp. 3311–3325, 1997.

[46] N. Caporale and Y. Dan, "Spike timing-dependent plasticity: a hebbian learning rule," *Annu. Rev. Neurosci.*, vol. 31, pp. 25–46, 2008.

[47] W. Lotter, G. Kreiman, and D. Cox, "Unsupervised learning of visual structure using predictive generative networks," *arXiv preprint arXiv:1511.06380*, 2015.

[48] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[49] M. Mathieu, C. Couprie, and Y. LeCun, "Deep multi-scale video prediction beyond mean square error," *arXiv preprint arXiv:1511.05440*, 2015.

[50] A. v. d. Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," *arXiv preprint arXiv:1601.06759*, 2016.

[51] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, "Building machines that learn and think like people," *Behavioral and Brain Sciences*, vol. 40, 2017.

[52] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, pp. 2672–2680, 2014.

[53] W. Gerstner, R. Kempter, J. L. van Hemmen, and H. Wagner, “A neuronal learning rule for sub-millisecond temporal coding,” *Nature*, vol. 383, pp. 76–78, 1996.

[54] R. Schultz, I. Gauthier, A. Klin, R. Fulbright, A. Anderson, F. Volkmar, P. Skudlarski, C. Lacadie, D. Cohen, and G. JC, “Abnormal ventral temporal cortical activity during face discrimination among individuals with autism and asperger syndrome,” *Arch Gen Psychiatry*, vol. 57, pp. 331–340, 2000.

[55] S. F. C. E. Pierce K, Haist F, “The brain response to personally familiar faces in autism: findings of fusiform activity and beyond,” *Brain*, vol. 127, pp. 2703–2716, 2004.

[56] N. Kriegeskorte, M. Mur, and P. Bandettini, “Representational similarity analysis–connecting the branches of systems neuroscience,” *Frontiers in Systems Neuroscience*, vol. 2, 2008.

[57] S.-M. Khaligh-Razavi and N. Kriegeskorte, “Deep supervised, but not unsupervised, models may explain IT cortical representation,” *PLoS Computational Biology*, vol. 10, no. 11, p. e1003915, 2014.

[58] U. Güçlü and M. A. van Gerven, “Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream,” *The Journal of Neuroscience*, vol. 35, no. 27, pp. 10005–10014, 2015.

[59] B. A. Wandell, S. O. Dumoulin, and A. A. Brewer, “Visual field maps in human cortex,” *Neuron*, vol. 56, no. 2, pp. 366–383, 2007.

[60] M. A. Silver and S. Kastner, “Topographic maps in human frontal and parietal cortex,” *Trends in Cognitive Sciences*, vol. 13, no. 11, pp. 488–495, 2009.

[61] L. G. Ungerleider and M. Mishkin, “Two cortical visual systems,” in *Analysis of Visual Behavior* (D. J. Ingle, M. A. Goodale, and R. J. W. Mansfield, eds.), Cambridge: MIT Press, 1982.

[62] M. A. Goodale and A. D. Milner, "Separate visual pathways for perception and action," *Trends in Neurosciences*, vol. 15, no. 1, pp. 20–25, 1992.

[63] D. H. Hubel and T. N. Wiesel, "Receptive fields of single neurones in the cat's striate cortex," *The Journal of Physiology*, vol. 148, no. 3, pp. 574–591, 1959.

[64] D. H. Hubel and T. N. Wiesel, "Receptive fields and functional architecture of monkey striate cortex," *The Journal of Physiology*, vol. 195, no. 1, pp. 215–243, 1968.

[65] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *The Journal of Physiology*, vol. 160, no. 1, pp. 106–154, 1962.

[66] G. C. DeAngelis, I. Ohzawa, and R. D. Freeman, "Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. I. General characteristics and postnatal development," *Journal of Neurophysiology*, vol. 69, no. 4, pp. 1091–1117, 1993.

[67] S. W. Kuffler, "Discharge patterns and functional organization of mammalian retina," *Journal of Neurophysiology*, vol. 16, no. 1, pp. 37–68, 1953.

[68] D. H. Hubel and T. N. Wiesel, "Integrative action in the cat's lateral geniculate body," *The Journal of Physiology*, vol. 155, no. 2, pp. 385–398, 1961.

[69] D. J. Heeger, "Normalization of cell responses in cat striate cortex," *Visual Neuroscience*, vol. 9, no. 02, pp. 181–197, 1992.

[70] J. L. Gardner, A. Anzai, I. Ohzawa, and R. D. Freeman, "Linear and nonlinear contributions to orientation tuning of simple cells in the cat's striate cortex," *Visual Neuroscience*, vol. 16, no. 06, pp. 1115–1121, 1999.

[71] A. Anzai, I. Ohzawa, and R. D. Freeman, "Neural mechanisms for processing binocular information II. Complex cells," *Journal of Neurophysiology*, vol. 82, no. 2, pp. 909–924, 1999.

[72] M. Carandini and D. J. Heeger, "Normalization as a canonical neural computation," *Nature Reviews Neuroscience*, vol. 13, no. 1, pp. 51–62, 2012.

[73] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological Cybernetics*, vol. 36, no. 4, pp. 193–202, 1980.

[74] Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. Howard, W. E. Hubbard, and L. D. Jackel, "Handwritten digit recognition with a back-propagation network," in *Advances in Neural Information Processing Systems*, pp. 396–404, 1990.

[75] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nature Neuroscience*, vol. 2, no. 11, pp. 1019–1025, 1999.

[76] N. Kanwisher, P. Downing, R. Epstein, Z. Kourtzi, R. Cabeza, and A. Kingstone, "Functional neuroimaging of visual cognition," *Handbook of Functional Neuroimaging of Cognition*, pp. 109–151, 2001.

[77] R. Malach, J. B. Reppas, R. R. Benson, K. K. Kwong, H. Jiang, W. A. Kennedy, P. J. Ledden, T. J. Brady, B. R. Rosen, and R. B. Tootell, "Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex," *Proceedings of the National Academy of Sciences*, vol. 92, no. 18, pp. 8135–8139, 1995.

[78] D. I. Perrett, E. T. Rolls, and W. Caan, "Visual neurones responsive to faces in the monkey temporal cortex," *Experimental Brain Research*, vol. 47, no. 3, pp. 329–342, 1982.

[79] R. Desimone, T. D. Albright, C. G. Gross, and C. Bruce, "Stimulus-selective properties of inferior temporal neurons in the macaque," *The Journal of Neuroscience*, vol. 4, no. 8, pp. 2051–2062, 1984.

[80] K. Tanaka, H.-a. Saito, Y. Fukada, and M. Moriya, "Coding visual images of objects in the inferotemporal cortex of the macaque monkey," *Journal of Neurophysiology*, vol. 66, no. 1, pp. 170–189, 1991.

[81] W. A. Freiwald and D. Y. Tsao, "Functional compartmentalization and viewpoint generalization within the macaque face-processing system," *Science*, vol. 330, no. 6005, pp. 845–851, 2010.

[82] J. W. Tanaka and M. J. Farah, "Parts and wholes in face recognition," *The Quarterly Journal of Experimental Psychology*, vol. 46, no. 2, pp. 225–245, 1993.

[83] N. Kanwisher and G. Yovel, "The fusiform face area: a cortical region specialized for the perception of faces," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 361, no. 1476, pp. 2109–2128, 2006.

[84] Y. Wada and T. Yamamoto, "Selective impairment of facial recognition due to a haematoma restricted to the right fusiform and lateral occipital region," *The Journal of Neurology, Neurosurgery & Psychiatry*, vol. 71, no. 2, pp. 254–257, 2001.

[85] S.-R. Afraz, R. Kiani, and H. Esteky, "Microstimulation of inferotemporal cortex influences face categorization," *Nature*, vol. 442, no. 7103, pp. 692–695, 2006.

[86] A. Afraz, E. S. Boyden, and J. J. DiCarlo, "Optogenetic and pharmacological suppression of spatial clusters of face neurons reveal their causal role in face gender discrimination," *Proceedings of the National Academy of Sciences*, vol. 112, no. 21, pp. 6730–6735, 2015.

[87] J. Parvizi, C. Jacques, B. L. Foster, N. Withoft, V. Rangarajan, K. S. Weiner, and K. Grill-Spector, "Electrical stimulation of human fusiform face-selective regions distorts face perception," *The Journal of Neuroscience*, vol. 32, no. 43, pp. 14915–14920, 2012.

[88] K. N. Kay, T. Naselaris, R. J. Prenger, and J. L. Gallant, "Identifying natural images from human brain activity," *Nature*, vol. 452, no. 7185, pp. 352–355, 2008.

[89] T. Naselaris, R. J. Prenger, K. N. Kay, M. Oliver, and J. L. Gallant, "Bayesian reconstruction of natural images from human brain activity," *Neuron*, vol. 63, no. 6, pp. 902–915, 2009.

[90] D. E. Stansbury, T. Naselaris, and J. L. Gallant, "Natural scene statistics account for the representation of scene categories in human visual cortex," *Neuron*, vol. 79, no. 5, pp. 1025–1034, 2013.

[91] J. L. Gardner, E. P. Merriam, J. A. Movshon, and D. J. Heeger, "Maps of visual space in human occipital cortex are retinotopic, not spatiotopic," *The Journal of Neuroscience*, vol. 28, no. 15, pp. 3988–3999, 2008.

[92] P. F. Van de Moortele, E. J. Auerbach, C. Olman, E. Yacoub, K. Ugurbil, and S. Moeller, "T1 weighted brain images at 7 Tesla unbiased for Proton Density, T2* contrast and RF coil receive B1 sensitivity with simultaneous vessel visualization," *Neuroimage*, vol. 46, no. 2, pp. 432–446, 2009.

[93] A. M. Dale, B. Fischl, and M. I. Sereno, "Cortical surface-based analysis: I. Segmentation and surface reconstruction," *Neuroimage*, vol. 9, no. 2, pp. 179–194, 1999.

[94] K. P. Pruessmann, M. Weiger, M. B. Scheidegger, and P. Boesiger, "SENSE: sensitivity encoding for fast MRI," *Magnetic Resonance in Medicine*, vol. 42, no. 5, pp. 952–962, 1999.

[95] X. Hu, T. H. Le, T. Parrish, and P. Erhard, "Retrospective estimation and correction of physiological fluctuation in functional MRI," *Magnetic Resonance in Medicine*, vol. 34, no. 2, pp. 201–212, 1995.

[96] P. Kellman, F. H. Epstein, and E. R. McVeigh, "Adaptive sensitivity encoding incorporating temporal filtering (TSENSE)," *Magnetic Resonance in Medicine*, vol. 45, no. 5, pp. 846–852, 2001.

[97] O. Nestares and D. J. Heeger, "Robust multiresolution alignment of MRI brain volumes," *Magnetic Resonance in Medicine*, vol. 43, no. 5, pp. 705–715, 2000.

[98] B. A. Wandell and J. Winawer, "Imaging retinotopic maps in the human brain," *Vision Research*, vol. 51, no. 7, pp. 718–737, 2011.

[99] D. Schluppeck, P. Glimcher, and D. J. Heeger, "Topographic organization for delayed saccades in human posterior parietal cortex," *Journal of Neurophysiology*, vol. 94, no. 2, pp. 1372–1384, 2005.

[100] J. D. Swisher, M. A. Halko, L. B. Merabet, S. A. McMains, and D. C. Somers, "Visual topography of human intraparietal sulcus," *The Journal of Neuroscience*, vol. 27, no. 20, pp. 5326–5337, 2007.

[101] R. Rajimehr, K. J. Devaney, N. Y. Bilenko, J. C. Young, and R. B. Tootell, "The "parahippocampal place area" responds preferentially to high spatial frequencies in humans and monkeys," *PLoS Biology*, vol. 9, no. 4, p. e1000608, 2011.

[102] N. Kanwisher, J. McDermott, and M. M. Chun, "The fusiform face area: a module in human extrastriate cortex specialized for face perception," *The Journal of Neuroscience*, vol. 17, no. 11, pp. 4302–4311, 1997.

[103] R. Epstein and N. Kanwisher, "A cortical representation of the local visual environment," *Nature*, vol. 392, no. 6676, pp. 598–601, 1998.

[104] J. Larsson and D. J. Heeger, "Two retinotopic visual areas in human lateral occipital cortex," *The Journal of Neuroscience*, vol. 26, no. 51, pp. 13128–13142, 2006.

[105] K. N. Kay, A. Rokem, J. Winawer, R. F. Dougherty, and B. A. Wandell, "GLMdenoise: a fast, automated technique for denoising task-based fMRI data," *Frontiers in Neuroscience*, vol. 7, 2013.

[106] T. M. Mitchell, S. V. Shinkareva, A. Carlson, K.-M. Chang, V. L. Malave, R. A. Mason, and M. A. Just, "Predicting human brain activity associated with the meanings of nouns," *Science*, vol. 320, no. 5880, pp. 1191–1195, 2008.

[107] J. Freeman and E. P. Simoncelli, "Metamers of the ventral stream," *Nature Neuroscience*, vol. 14, no. 9, pp. 1195–1201, 2011.

[108] T. Serre, A. Oliva, and T. Poggio, "A feedforward architecture accounts for rapid categorization," *Proceedings of the National Academy of Sciences*, vol. 104, no. 15, pp. 6424–6429, 2007.

[109] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, "A limited memory algorithm for bound constrained optimization," *SIAM Journal on Scientific Computing*, vol. 16, no. 5, pp. 1190–1208, 1995.

[110] J. H. Steiger, "Tests for comparing elements of a correlation matrix.," *Psychological Bulletin*, vol. 87, no. 2, pp. 245–251, 1980.

[111] I. A. Lee and K. J. Preacher, "Calculation for the Test of the Difference Between Two Dependent Correlations with One Variable in Common." http://quantpsy.org/corrtest/corrtest2.htm, 2013. Accessed: 2014-04-01.

[112] H. Nili, C. Wingfield, A. Walther, L. Su, W. Marslen-Wilson, and N. Kriegeskorte, "A toolbox for representational similarity analysis," *PLoS Computational Biology*, vol. 10, no. 4, p. e1003553, 2014.

[113] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition*, pp. 248–255, IEEE, 2009.

[114] S. O. Dumoulin and B. A. Wandell, "Population receptive field estimates in human visual cortex," *Neuroimage*, vol. 39, no. 2, pp. 647–660, 2008.

[115] C. F. Cadieu, H. Hong, D. L. Yamins, N. Pinto, D. Ardila, E. A. Solomon, N. J. Majaj, and J. J. DiCarlo, "Deep neural networks rival the representation

of primate IT cortex for core visual object recognition," *PLoS Computational Biology*, vol. 10, no. 12, p. e1003963, 2014.

[116] M. Ito, H. Tamura, I. Fujita, and K. Tanaka, "Size and position invariance of neuronal responses in monkey inferotemporal cortex," *Journal of Neurophysiology*, vol. 73, no. 1, pp. 218–226, 1995.

[117] J. J. DiCarlo, D. Zoccolan, and N. C. Rust, "How does the brain solve visual object recognition?," *Neuron*, vol. 73, no. 3, pp. 415–34, 2012.

[118] D. J. Felleman and D. C. Van Essen, "Distributed hierarchical processing in the primate cerebral cortex," *Cerebral Cortex*, vol. 1, no. 1, pp. 1–47, 1991.

[119] A. Pasupathy and C. E. Connor, "Shape representation in area V4: position-specific tuning for boundary conformation," *Journal of Neurophysiology*, vol. 86, no. 5, pp. 2505–2519, 2001.

[120] A. Pasupathy and C. E. Connor, "Population coding of shape in area V4," *Nature Neuroscience*, vol. 5, no. 12, pp. 1332–1338, 2002.

[121] M. Ito and H. Komatsu, "Representation of angles embedded within contour stimuli in area V2 of macaque monkeys," *The Journal of Neuroscience*, vol. 24, no. 13, pp. 3313–3324, 2004.

[122] J. L. Gallant, J. Braun, and D. C. Van Essen, "Selectivity for polar, hyperbolic, and cartesian gratings in macaque visual cortex," *Science*, vol. 259, no. 5091, pp. 100–103, 1993.

[123] J. Hegde and D. C. Van Essen, "Temporal dynamics of 2D and 3D shape representation in macaque visual area V4," *Visual Neuroscience*, vol. 23, no. 05, pp. 749–763, 2006.

[124] A. Anzai, X. Peng, and D. C. Van Essen, "Neurons in monkey visual area V2 encode combinations of orientations," *Nature Neuroscience*, vol. 10, no. 10, pp. 1313–1321, 2007.

[125] J. L. Gallant, C. E. Connor, S. Rakshit, J. W. Lewis, and D. C. Van Essen, "Neural responses to polar, hyperbolic, and Cartesian gratings in area V4 of the macaque monkey," *Journal of Neurophysiology*, vol. 76, no. 4, pp. 2718–2739, 1996.

[126] J. Freeman, C. M. Ziemba, D. J. Heeger, E. P. Simoncelli, and J. A. Movshon, "A functional and perceptual signature of the second visual area in primates," *Nature Neuroscience*, vol. 16, no. 7, pp. 974–981, 2013.

[127] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision*, pp. 818–833, Springer, 2014.

[128] M. Carandini, D. J. Heeger, and J. A. Movshon, "Linearity and normalization in simple cells of the macaque primary visual cortex," *The Journal of Neuroscience*, vol. 17, no. 21, pp. 8621–8644, 1997.

[129] J. K. Steeves, J. C. Culham, B. C. Duchaine, C. C. Pratesi, K. F. Valyear, I. Schindler, G. K. Humphrey, A. D. Milner, and M. A. Goodale, "The fusiform face area is not sufficient for face recognition: evidence from a patient with dense prosopagnosia and no occipital face area," *Neuropsychologia*, vol. 44, no. 4, pp. 594–609, 2006.

[130] T. Konkle and A. Oliva, "A real-world size organization of object responses in occipitotemporal cortex," *Neuron*, vol. 74, no. 6, pp. 1114–1124, 2012.

[131] E. A. Murray, T. J. Bussey, and L. M. Saksida, "Visual Perception and Memory: A New View of Medial Temporal Lobe Function in Primates and Rodents," *Annual Review of Neuroscience*, vol. 30, pp. 99–122, 2007.

[132] J. L. Gardner, P. Sun, R. A. Waggoner, K. Ueno, K. Tanaka, and K. Cheng, "Contrast adaptation and representation in human early visual cortex," *Neuron*, vol. 47, no. 4, pp. 607–620, 2005.

[133] M. Costagli, K. Ueno, P. Sun, J. L. Gardner, X. Wan, E. Ricciardi, P. Pietrini, K. Tanaka, and K. Cheng, "Functional signalers of changes in visual stimuli: cortical responses to increments and decrements in motion coherence," *Cerebral Cortex*, vol. 24, no. 1, pp. 110–118, 2014.

[134] R. K. Yin, "Looking at upside-down faces," *Journal of Experimental Psychology*, vol. 81, no. 1, p. 141, 1969.

[135] E. H. De Haan, A. Young, and F. Newcombe, "Faces interfere with name classification in a prosopagnosic patient," *Cortex*, vol. 23, no. 2, pp. 309–316, 1987.

[136] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv.org*, vol. abs/1312.6199, 2013.

[137] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," *arXiv.org*, vol. abs/1412.1897, 2014.

[138] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, pp. 255–258, 1995.

[139] G. E. Hinton and R. S. Zemel, "Autoencoders, minimum description length, and helmholtz free energy.," *Advances in Neural Information Processing Systems*, 1994.

[140] K. A. Coates, A. and A. Y. Ng, "Emergence of object-selective features in unsupervised feature learning," *Advances in Neural Information Processing Systems*, pp. 2681–2689, 2012.

[141] Q. V. Le, "Building high-level features using large scale unsupervised learning," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8595–8598, IEEE, 2013.

[142] P. Vincent, H. Larochelle, Y. Bengio, and P. A. Manzagol, "Extracting and composing robust features with denoising autoencoders," *Proceedings of the 25th international conference on Machine learning*, pp. 1096–1103, 2008.

[143] Y. Bengio, "Learning deep architectures for ai," *Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[144] C. A. Bengio, Y. and P. Vincent, "Representation learning: A review and new perspectives," *Pattern Analysis and Machine Intelligence, IEEE Transactions*, vol. 35, no. 8, pp. 1798–1828, 2013.

[145] J. Hurri, A. Hyvarinen, and E. Oja, "Wavelets and natural image statistics," *Proceedings of the 10th Scandinavian Conference on Image Analysis, Pattern Recognition Society of Finland*, pp. 13—18, 1997.

[146] C. Kanan and G. Cottrell, "Robust classification of objects, faces, and flowers using natural image statistics," *Computer Vision and Pattern Recognition*, pp. 2472—2479, 2010.

[147] E. P. Simoncelli and B. A. Olshausen, "Natural image statistics and neural representation," *Annual Review of Neuroscience*, vol. 24, no. 1, pp. 1193–1216, 2001.

[148] J. Bruna and S. Mallat, "Invariant scattering convolution network," *IEEE Trans. on PAMI*, vol. 35, no. 8, pp. 1872—1886, 2013.

[149] A. Torralba and A. Oliva, "Statistics of natural image categories," *Network: Computation in Neural Systems*, pp. 391—412, 2003.

[150] N. de Freitas, "Predicting parameters in deep learning," 2013.

[151] R. N. Bracewell, *The Fourier transform and its applications.* 1986.

[152] Krizhevzky, A, "cuda-convnet - High-performance C++/CUDA implementation of convolutional neural networks," 2014.

[153] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint*, 2014.

[154] A. M. Saxe, J. L. McClelland, and S. Ganguli, "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks," *arXiv*, 2013.

[155] J. J. DiCarlo and D. D. Cox, "Untangling invariant object recognition," *Trends in Cognitive Sciences*, vol. 11, no. 8, pp. 333–341, 2007.

[156] K. Grill-Spector, Z. Kourtzi, and N. Kanwisher, "The lateral occipital complex and its role in object recognition," *Vision Research*, vol. 41, no. 10-11, pp. 1409–1422, 2001.

[157] R. Malach, I. Levy, and U. Hasson, "The topography of high-order human object areas," *Trends in Cognitive Sciences*, vol. 6, no. 4, pp. 176–184, 2002.

[158] N. Kriegeskorte, M. Mur, D. A. Ruff, R. Kiani, J. Bodurka, H. Esteky, K. Tanaka, and P. A. Bandettini, "Matching categorical object representations in inferior temporal cortex of man and monkey," *Neuron*, vol. 60, no. 6, pp. 1126–41, 2008.

[159] N. Kriegeskorte, "Relating population-code representations between man, monkey, and computational models," *Frontiers in Neuroscience*, vol. 3, p. 35, 2009.

[160] K. Tanaka, "Inferotemporal cortex and object vision," *Annual Review of Neuroscience*, vol. 19, pp. 109–139, 1996.

[161] N. K. Logothetis and D. L. Sheinberg, "Visual object recognition," *Annual Review of Neuroscience*, vol. 19, no. 1, pp. 577–621, 1996.

[162] C. G. Gross, "How inferior temporal cortex became a visual area," *Cerebral Cortex*, vol. 4, no. 5, pp. 455–469, 1994.

[163] R. Vogels and G. Orban, "Activity of inferior temporal neurons during orientation discrimination with successively presented gratings," *Journal of Neurophysiology*, vol. 71, pp. 1428–1451, 1994.

[164] C. E. Connor, S. L. Brincat, and A. Pasupathy, "Transformation of shape information in the ventral pathway," *Current Opinion in Neurobiology*, vol. 17, no. 2, pp. 140–147, 2007.

[165] D. H. Hubel, T. N. Wiesel, and S. LeVay, "Plasticity of ocular dominance columns in monkey striate cortex," *Phil. Trans. R. Soc. Lond. B*, vol. 278, no. 961, pp. 377–409, 1977.

[166] R. O. Wong, M. Meister, and C. J. Shatz, "Transient period of correlated bursting activity during development of the mammalian retina," *Neuron*, vol. 11, no. 5, pp. 923–938, 1993.

[167] N. J. Priebe and D. Ferster, "Mechanisms of neuronal computation in mammalian visual cortex," *Neuron*, vol. 75, no. 2, pp. 194–208, 2012.

[168] H. R. Rodman, S. Scalaidhe, and C. G. Gross, "Response properties of neurons in temporal cortical visual areas of infant monkeys," *Journal of Neurophysiology*, vol. 70, no. 3, pp. 1115–1136, 1993.

[169] R. Baillargeon, E. S. Spelke, and S. Wasserman, "Object permanence in five-month-old infants," *Cognition*, vol. 20, no. 3, pp. 191–208, 1985.

[170] J. Piaget, *The construction of reality in the child*, vol. 82. Routledge, 1954.

[171] T. N. Wiesel and D. H. Hubel, "Single-cell responses in striate cortex of kittens deprived of vision in one eye," *Journal of neurophysiology*, vol. 26, no. 6, pp. 1003–1017, 1963.

[172] J. A. Movshon and R. C. Van Sluyters, "Visual neural development," *Annual review of psychology*, vol. 32, no. 1, pp. 477–522, 1981.

[173] K. A. Stavros and L. Kiorpes, "Behavioral measurement of temporal contrast sensitivity development in macaque monkeys (macaca nemestrina)," *Vision research*, vol. 48, no. 11, pp. 1335–1344, 2008.

[174] L. Kiorpes and S. A. Bassin, "Development of contour integration in macaque monkeys," *Visual neuroscience*, vol. 20, no. 5, pp. 567–575, 2003.

[175] C. Hall-Haro and L. Kiorpes, "Normal development of pattern motion sensitivity in macaque monkeys," *Visual neuroscience*, vol. 25, no. 5-6, pp. 675–684, 2008.

[176] L. Kiorpes, T. Price, C. Hall-Haro, and J. A. Movshon, "Development of sensitivity to global form and motion in macaque monkeys (macaca nemestrina)," *Vision research*, vol. 63, pp. 34–42, 2012.

[177] L. Kiorpes, "Visual development in primates: neural mechanisms and critical periods," *Developmental neurobiology*, vol. 75, no. 10, pp. 1080–1090, 2015.

[178] S. Amemori, K.-i. Amemori, M. L. Cantor, and A. M. Graybiel, "A non-invasive head-holding device for chronic neural recordings in awake behaving monkeys," *The Journal of Neuroscience Methods*, vol. 240, pp. 154–160, 2015.

[179] R. Rajalingham, E. B. Issa, P. Bashivan, K. Kar, K. Schmidt, and J. J. DiCarlo, "Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks," *bioRxiv*, p. 240614, 2018.

[180] E. T. Rolls and M. J. Tovee, "Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex," *Journal of neurophysiology*, vol. 73, no. 2, pp. 713–726, 1995.

[181] R. Rajalingham, K. Schmidt, and J. J. DiCarlo, "Comparison of object recognition behavior in human and monkey," *The Journal of Neuroscience*, vol. 35, no. 35, pp. 12127–12136, 2015.

[182] F. O. Galashan, H. C. Rempel, A. Meyer, E. Gruber-Dujardin, A. K. Kreiter, and D. Wegener, "A new type of recording chamber with an easy-to-exchange microdrive array for chronic recordings in macaque monkeys," *Journal of Neurophysiology*, vol. 105, no. 6, pp. 3092–3105, 2011.

[183] E. B. Issa, A. M. Papanastassiou, and J. J. DiCarlo, "Large-scale, high-resolution neurophysiological maps underlying fMRI of macaque temporal lobe," *The Journal of Neuroscience*, vol. 33, no. 38, pp. 15207–15219, 2013.

[184] A.-M. M. Oswald and A. D. Reyes, "Maturation of intrinsic and synaptic properties of layer 2/3 pyramidal neurons in mouse auditory cortex," *Journal of Neurophysiology*, vol. 99, no. 6, pp. 2998–3008, 2008.

[185] M. L. Belleau and R. A. Warren, "Postnatal development of electrophysiological properties of nucleus accumbens neurons," *Journal of Neurophysiology*, vol. 84, no. 5, pp. 2204–2216, 2000.

[186] A. S. Ramoa and D. A. McCormick, "Developmental changes in electrophysiological properties of lgnd neurons during reorganization of retinogeniculate connections," *The Journal of Neuroscience*, vol. 14, no. 4, pp. 2089–2097, 1994.

[187] S. G. Waxman, "Determinants of conduction velocity in myelinated nerve fibers," *Muscle & nerve*, vol. 3, no. 2, pp. 141–150, 1980.

[188] C. M. Ziemba, J. Freeman, J. A. Movshon, and E. P. Simoncelli, "Selectivity and tolerance for visual texture in macaque V2," *Proceedings of the National Academy of Sciences*, vol. 113, no. 22, pp. E3140–E3149, 2016.

[189] R. Desimone, T. D. Albright, C. G. Gross, and C. Bruce, "Stimulus-selective properties of inferior temporal neurons in the macaque," *The Journal of Neuroscience*, vol. 4, no. 8, pp. 2051–62, 1984.

[190] T. N. Chandrapala, *Sparsity and Temporal Slowness as Principles Underlying the Development of Visual Receptive Fields.* Hong Kong University of Science and Technology (Hong Kong), 2017.

[191] W. Böhmer, S. Grünewälder, H. Nickisch, and K. Obermayer, "Regularized sparse kernel slow feature analysis," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 235–248, Springer, 2011.

[192] K.-C. Hsieh, E. L. Robinson, and C. A. Fuller, "Sleep architecture in unrestrained rhesus monkeys (macaca mulatta) synchronized to 24-hour light-dark cycles," *Sleep*, vol. 31, no. 9, pp. 1239–1250, 2008.

[193] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell, "Curiosity-driven exploration by self-supervised prediction," in *International Conference on Machine Learning (ICML)*, vol. 2017, 2017.

[194] M. Bellemare, S. Srinivasan, G. Ostrovski, T. Schaul, D. Saxton, and R. Munos, "Unifying count-based exploration and intrinsic motivation," in *Advances in Neural Information Processing Systems*, pp. 1471–1479, 2016.